Universidad Carlos III de Madrid
Facultad de Humanidades, Comunicación y Documentación
Departamento de Biblioteconomía y Documentación

Doctorado en Documentación
Archivos y Bibliotecas en el Entorno Digital

ANÁLISIS DE LOS CRITERIOS DE RELEVANCIA DOCUMENTAL MEDIANTE CONSULTAS DE INFORMACIÓN EN EL ENTORNO WEB

Tesis Doctoral

Autor
Valentín Moreno Pelayo

Director
Prof. Dr. D. José Antonio Moreiro González
Prof. Dr. Dña. Sonia Sánchez Cuadrado

Getafe, julio de 2010

TESIS DOCTORAL

ANÁLISIS DE LOS CRITERIOS DE RELEVANCIA DOCUMENTAL MEDIANTE CONSULTAS DE INFORMACIÓN EN EL ENTORNO WEB

Autor: Valentín Moreno Pelayo

Director/es: José Antonio Moreiro González
Sonia Sánchez Cuadrado

Firma del Tribunal Calificador:

Firma

Presidente: (Nombre y apellidos)

Vocal: (Nombre y apellidos)

Vocal: (Nombre y apellidos)

Vocal: (Nombre y apellidos)

Secretario: (Nombre y apellidos)

Calificación:

Leganés/Getafe, de de

A Manoli y mi familia

Resumen

La búsqueda de información no se entiende sin los motores de búsqueda web. Ante una demanda de información los buscadores web ordenan los resultados de forma que las páginas web más relevantes para la consulta aparezcan en las primeras posiciones. Esto genera un alto grado de competitividad entre las páginas web por obtener mejores asignaciones de relevancia por parte de los buscadores. Por norma general, los usuarios suelen consultar sólo los primeros resultados que devuelve un motor de búsqueda, en consecuencia ocupar estos puestos se traduce en mayor prestigio y visibilidad. Por tanto, la percepción de relevancia documental web por parte de los usuarios está intrínsecamente unida a los motores de búsqueda.

En este trabajo se propone y desarrolla una metodología para determinar la relevancia documental web de forma automática, que se puede interpretar como: predicción automática de la posición que otorgaría un motor de búsqueda a un documento web entre los resultados de una consulta.

La investigación se completa identificando los factores considerados en el posicionamiento web, a partir del estudio de herramientas empleadas en la optimización y promoción de páginas web. También se analiza el peso de cada uno de estos factores en los algoritmos de ordenación de los buscadores.

Finalmente, en relación a las capacidades adquiridas para emular el comportamiento de los motores de búsqueda se propone un método de optimización web que estima previamente la rentabilidad del proceso. De esta forma no se invertirá en una campaña de promoción si los pronósticos de mejora del posicionamiento no se juzgan adecuados.

Agradecimientos

Al escribir estas líneas pienso en las personas que me han acompañado desde los orígenes de este trabajo y me siento afortunado. He recibido ayuda, consejos, sosiego y apoyo moral de cada uno de ellos según la faceta que ocupan en mi vida.

Aunque de forma escueta, una cualidad que me caracteriza, me gustaría expresar mi gratitud a:

- Jorge por mostrarme en sus clases lo que ahora es mi tema de investigación.
- Mis directores de tesis por su buen hacer profesional, la disponibilidad mostrada en todo momento, que ha permitido que pueda culminar esta investigación.
- Todos los miembros del grupo de investigación *Knowledge Reuse* y al grupo de desarrollo del Profesor Dr. Juan Llorens, por su calidad humana, por ser unos excelentes compañeros y amigos, y por su afán colaborativo. Todos ellos me hacen disfrutar de este trabajo día tras día.
- Mi familia y a todos aquellos que los siento como tales, porque me refuerzan y estimulan en todas las situaciones de la vida.
- Mis amigos, algunos ya incluidos en los agradecimientos anteriores, en permanente demostración de su condición especialmente en los momentos de adversidad.
- Y como no, a mi compañera en el más amplio sentido de la palabra.

Tabla de Contenidos

Índice de Figuras

Índice de Tablas

Capítulo I: Introducción

1.1 Motivación

Dentro de la Recuperación de Información Web, los motores de búsqueda de Internet juegan un papel fundamental en la sociedad. Millones de documentos están disponibles en la Web y los internautas acceden a ellos a través de los motores de búsqueda, ya sean genéricos o especializados.

Con cada consulta que se efectúa a un motor de búsqueda se obtienen miles o millones de resultados que son ordenados mediante algoritmos de posicionamiento web. Estos algoritmos de posicionamiento web están basados en múltiples factores y criterios de relevancia documental extraídos de la información de los documentos.

Es bien conocido que los usuarios comunes no consultan más de unos 10 ó 20 resultados devueltos por un motor de búsqueda (Silverstein et al., 1998), por lo que la posición que consiga una página o un documento ante determinada consulta es determinante para que sea visible y se conozca determinado sitio web. Esto coloca a los motores de búsqueda en una posición de poder y de responsabilidad social (Introna y Nissenbaum, 2000).

La recuperación de un documento en la Web está condicionada por los algoritmos de posicionamiento web. Estos algoritmos son secretos comerciales (Arasu et al., 2001) y varían periódicamente para evitar manipulaciones.

Tal es la importancia de que los sitios web o los documentos de la Web queden posicionados en las primeras posiciones del ranking (Ferber, 2003) que se han desarrollado múltiples iniciativas de optimización web. La optimización web es la aplicación de técnicas que mejoren los valores de los criterios tenidos en cuenta por los algoritmos de posicionamiento de los motores web.

Existen diferentes recomendaciones y propuestas software sobre la optimización web tanto gratuitas como comerciales, y sin embargo ninguna permite estimar el efecto que tendrá en el ranking de posiciones las modificaciones y mejoras que recomiendan las herramientas SEO (Search Engine Optimization).

El problema es que se invierte en campañas de promoción sin saber si finalmente serán rentables. Incluso grandes progresos en el posicionamiento de una web puede que no sean significativos en cuanto al aumento del volumen de negocio que publicita. Por ejemplo, una página que ascienda de la posición 500 a la 100 en los resultados de una consulta seguirá siendo casi invisible para los usuarios (Silverstein et al., 1998).

1.2 Hipótesis

El análisis y la comparación de los criterios de relevancia en las herramientas SEO y en los algoritmos de posicionamiento ante una consulta dada en un motor determinado, permitirá establecer un método para la estimación de la relevancia y la estimación de la posición que alcanzará un documento ante una consulta en determinado motor.

1.3 Objeto

Proponer un método para la estimación de la relevancia de un documento para una consulta determinada respecto a los documentos de su competencia. El método se centra en el posicionamiento de documentación web ante consultas concretas para cualquier motor de búsqueda.

La principal aplicación práctica es predecir las mejoras conseguidas antes de realizar los procesos de optimización con las herramientas SEO.

1.4 Objetivos

La presente propuesta plantea estimar la relevancia documental para una consulta en los motores de búsqueda web respecto a los documentos de su competencia. Para ello se definen los siguientes objetivos:

- Identificar los factores que los motores de búsqueda consideran más influyentes en el posicionamiento.
- Estimar automáticamente la relevancia documental asignada por los buscadores web.

- Determinar de forma automática el grado de influencia de cada factor en el posicionamiento efectuado por los motores de búsqueda.

En la propuesta de optimización de una web se modificarán los valores de sus características o atributos tal como si se hubiese llevado a cabo el proceso de promoción, pero sin realizarlo. Se estima con un modelo predictivo la posición que ocuparía esa web entre su competencia con esos valores hipotéticos, y en caso de que se considere rentable, se efectúan realmente las mejoras. En caso contrario se puede intentar otra propuesta de optimización (Figura I-1).

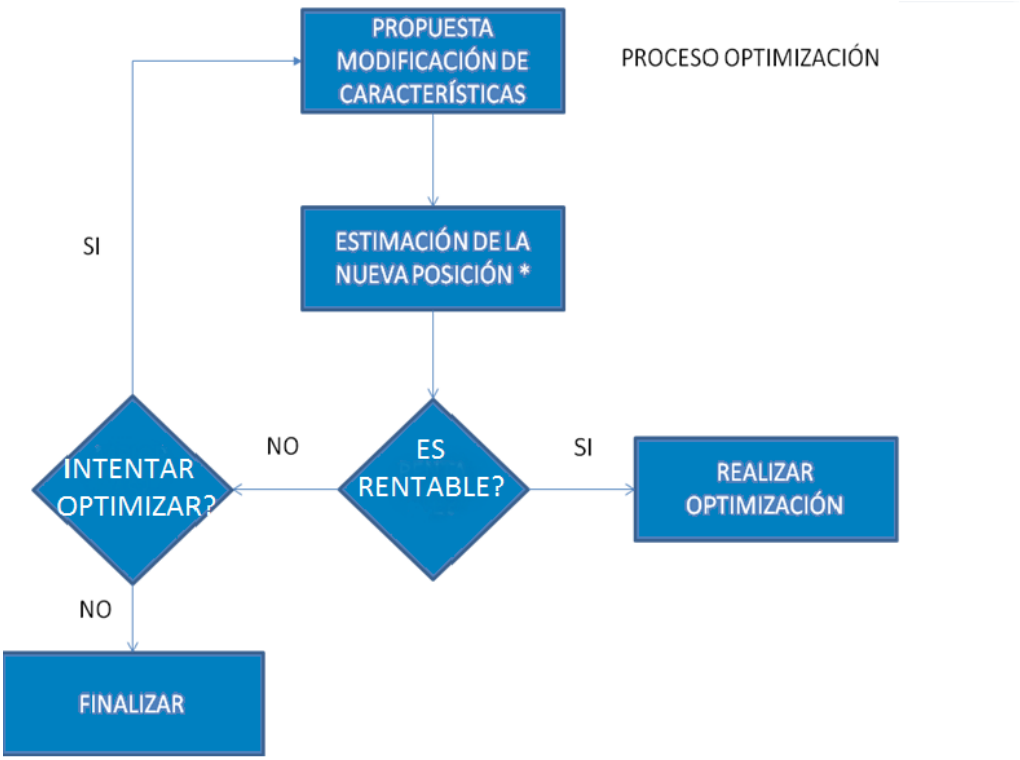


Figura I-1: Principales pasos de la predicción de mejoras de optimización

Un supuesto sencillo. Si se quisiera enlazar una página a optimizar por otra página de PageRank 5 mediante pago de 1000€ (precio acorde con (López, 2009)) se sumaría un enlace más al atributo "*enlaces entrantes*", y se predeciría la nueva posición alcanzada con este cambio. Sólo en el caso de que la posición estimada sea satisfactoria se realizaría la optimización en este caso por compra, si no se podrían proponer otras alternativas o desistir.

1.5 Metodología

Para cumplir los objetivos se han establecido las siguientes fases en la metodología:

i. Detección de los factores relevantes del posicionamiento web utilizados por las herramientas SEO.

ii. Estimación de la relevancia documental asignada por los buscadores web.

iii. Determinación automática de los factores más influyentes en el posicionamiento ejercido por los motores de búsqueda.

El desarrollo de la investigación se estructura en varios bloques que permiten cumplir el objetivo principal propuesto en la tesis. Los bloques 2 y 3 comparten la estructura, mientras que el bloque 1 tiene un desarrollo diferente (Figura I-2).

En el bloque 1 se realizará un estudio de los factores de posicionamiento que se evalúan en las herramientas SEO con el objetivo de determinar cuáles de esos factores tienen mayor presencia en las herramientas y son más relevantes en el posicionamiento web. Este bloque se estructura en las siguientes fases:1) metodología; 2) resultados de la evaluación de las herramientas; 3) conclusiones obtenidas y discusión.

Los bloques 2 y 3 están constituidos por las fases de: metodología, experimentación, evaluación y resultados, discusión y conclusiones.

- En la **fase de metodología** se expone, sin llegar al nivel de detalle, los pasos a seguir en la investigación para alcanzar el objetivo propuesto.
- En la **desarrollo** se explica cada uno de los pasos expuestos en la metodología concretando cualquier particularidad relevante e incluyendo las decisiones que haya sido necesario adoptar.
- En la **experimentación y resultados** se muestran cada uno de los experimentos y sus resultados correspondientes, incluyendo si es el caso los de evaluación.
- En la **discusión y conclusiones** se presentan las conclusiones y los razonamientos que han conducido a ellas.

Figura I-2: Principales pasos de la metodología

1.6 Requisitos de la Investigación

En este punto se describen los requisitos de la investigación de acuerdo a los aspectos que delimitan el ámbito de estudio. Este estudio está restringido por requisitos temporales y por requisitos de las aplicaciones de software libre.

Las implicaciones temporales de los procesos de investigación son en este campo determinantes debido al fuerte dinamismo de las tecnologías web. Los datos recogidos para la investigación se obtuvieron en el periodo (2005-2008), no obstante el proceso metodológico es adaptable a los posibles cambios de los algoritmos de los sistemas de recuperación. Por consiguiente debe tenerse en cuenta los periodos en los que se han realizado los análisis y los experimentos ya que algunas situaciones específicas habrán cambiado, aunque con la vigencia de las aportaciones metodológicas.

Por otra parte, las recopilaciones automáticas de datos procedentes de la Web deben realizarse en un corto intervalo de tiempo, de lo contrario el permanente cambio del entorno web, podría invalidar las conclusiones obtenidas de los experimentos. Por ejemplo, no se pueden extraer conclusiones del funcionamiento de un algoritmo de ordenación con datos recogidos en periodos de tiempo diferentes, ya que es posible que el algoritmo hubiese evolucionado antes de finalizar la captura de esos datos.

Se ha priorizado las aplicaciones de software libre para los recursos de la investigación dado que no existe una financiación adicional que permita la compra de licencias de software. El funcionamiento de las aplicaciones incorporadas está garantizado por su larga trayectoria en investigación, como la herramienta de análisis de datos Weka (Witten y Frank, 2005). Otras, sin embargo, suelen ofrecer funcionalidades gratuitas, por ejemplo de captura de datos, que luego protegen mediante *captchas* con el posible fin de obtener beneficios por su uso automatizado. Esto ha excluido el análisis de los factores de posicionamiento en los que no había garantía de gratuidad.

Algunas de las aplicaciones de software relacionadas con la extracción de los datos han mostrado limitaciones en el paralelismo de sus funcionalidades. Además, las aplicaciones están especializadas en páginas web con formato HTML, ya que una amplia mayoría de las páginas web están en este formato.

Capítulo II: Estado del Arte

2.1 El Posicionamiento Web

2.1.1 Introducción

El posicionamiento web consiste en la ordenación de recursos web por su idoneidad como respuesta ante una búsqueda de información. Desde otro punto de vista, el posicionamiento web es el conjunto de criterios con los que los buscadores web asignan la relevancia a las páginas web asociadas a una consulta.

El posicionamiento web según los estudios realizados por Lluis Codina (2004) puede clasificarse en dos tipos de posicionamiento: posicionamiento natural y posicionamiento planificado.

El **Posicionamiento natural** es la relevancia que obtiene un sitio web en el que no se ha tenido en cuenta una planificación previa orientada a mejorar su visibilidad. Sin embargo, los contenidos apropiados de un recurso web junto a un diseño amigable no garantizan que alcance un buen posicionamiento.

El **Posicionamiento planificado** es el posicionamiento que obtiene un recurso web tras una planificación dirigida a mejorar la relevancia que le otorgarán los motores de búsqueda. Las tareas de optimización de su posicionamiento deben de considerarse desde los inicios del proyecto web, de lo contrario, los esfuerzos posteriores encaminados a alcanzar mejores posiciones tendrán gran repercusión en los costes.

El posicionamiento planificado ofrece dos vertientes:(1) la planificación fraudulenta en la que se pretende engañar al buscador sobre la temática y/o la relevancia de un documento web; (2) y la planificación ética en la que se tiene en cuenta el conocimiento disponible

sobre un motor de búsqueda con el fin de no perder ventaja competitiva frente a otros recursos de inferior calidad o de diferente temática (Glöggler, 2003).

Los buscadores deben estar siempre alerta ante estrategias de planificaciones fraudulentas, si un gran número de recursos web se posicionarán con estos métodos la calidad de los resultados que ofrece un motor de búsqueda quedaría mermada y por tanto, los usuarios perderían la confianza en el mismo y dejarían probablemente de utilizarlo.

2.1.2 Gestión del conocimiento Web

Son diversos los motivos implicados en la importancia de la gestión del conocimiento y por tanto responsables de la alta atención investigadora que se le dedica. Fensel (Fensel et al., 2002) destacó que la importancia de la Gestión del Conocimiento se fundamenta, principalmente, en los siguientes factores: la sobrecarga de información, la necesidad de recuperar información mediante consultas eficientes que no se limiten a las palabras clave, la falta de autoridad literaria en entornos web y la carencia de sistemas automáticos de procesamiento del lenguaje natural especializados en la gestión del conocimiento. De forma análoga, Antoniou y Harmelen (2004) indican que esta necesidad viene impulsada por las organizaciones, los colectivos y distintas disciplinas , con el fin de estructurar la información y así apoyar los procesos comunicativos, competitivos y de desarrollo interno.

En la actualidad, resulta evidente que el desarrollo de Internet ha provocado un aumento desmesurado de la información. Este mismo fenómeno puede ser observado en otras redes y organizaciones. Sin la gestión adecuada de ese conocimiento resulta imposible no sólo su recuperación sino incluso conocer la existencia de un recurso (Choi et al., 2003).

2.1.3 Recuperación de Información Web

El dinamismo de las colecciones documentales y de las consultas que se envían al sistema determina el modo en que se trata la recuperación.

El enfoque de recuperación adhoc (Türker, 2004) está especializado en los sistemas de Recuperación de Información (IR) tradicionales que se emplean sobre colecciones de documentos estables, es decir, colecciones en las que casi todos sus documentos permanecen inalterables mientras se realizan las consultas. En contraposición el enfoque de filtrado se utiliza en colecciones dinámicas con solicitudes estables de información. En

este enfoque, la principal tarea no es la clasificación de documentos, sino la creación de perfiles que representen las preferencias del usuario. Ejemplos de estas dos alternativas de recuperación son respectivamente: la recuperación en bibliotecas digitales y los sistemas información meteorológica.

Los sistemas IR convencionales hacen uso de índices para recuperar los documentos. Las consultas se basan en palabras clave que, al mostrarse descontextualizadas, sufren, al igual que los índices, una pérdida de su significado implícito (Baeza-Yates y Ribeiro-Neto, 1999).

Los términos para la indexación suelen ser sustantivos debido a su mayor carga semántica. Otras categorías gramaticales como los adjetivos son menos útiles debido a su carácter complementario. La importancia de los índices depende de su frecuencia de aparición en la colección de documentos. Así, las palabras presentes en todos los documentos son de poca utilidad debido a su nulo poder discriminante. Sin embargo, las palabras más infrecuentes en el conjunto de documentos tienen mayor potencial para reducir el número de documentos recuperados (Türker, 2004).

Existen diversos modelos de IR. En la siguiente figura se muestra una clasificación de estos modelos IR según los autores Baeza-Yates y Ribeiro-Neto (1999).

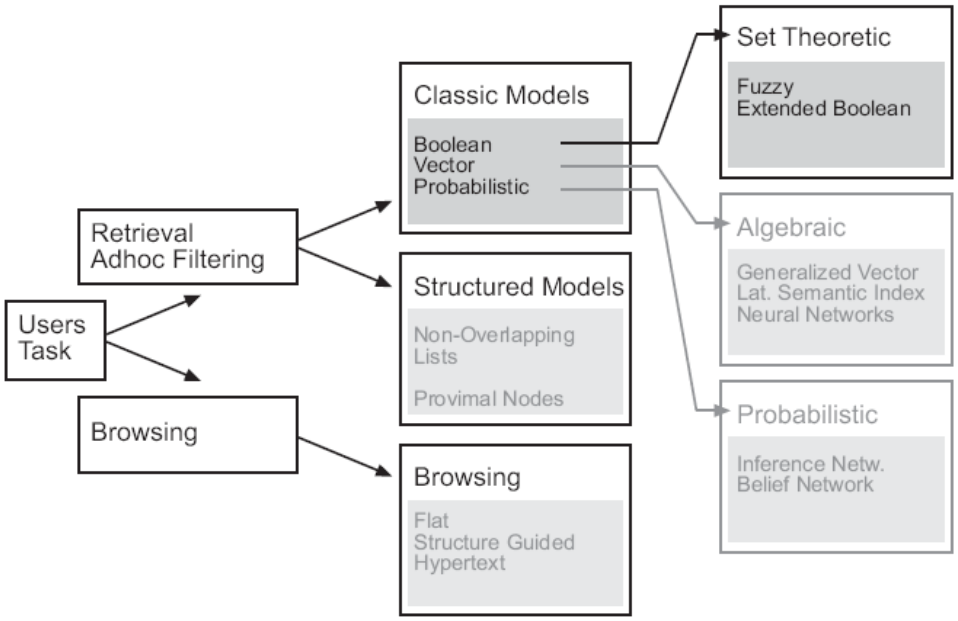


Figura II-1: Modelos de recuperación de información según Baeza-Yates & Ribero-Neto 1999 (p. 21)

El primer modelo clásico utilizado en IR es el **modelo booleano** (Rijsbergen, 1979), los índices en este modelo tienen asociados pesos binarios (0, 1) y se relacionan mediante operadores booleanos (Moens, 2000). Por tanto, se puede aplicar un posible tipo de búsqueda por operador (AND, OR, NOT) (Chowdhury y Chowdhury, 2001). En este modelo no hay coincidencia parcial con las consultas clasificándose cada documento como pertinente o no pertinente. Esta rigidez puede conducir a recuperaciones con alto grado de ruido o de silencio. Por último indicar que sólo cuando los usuarios son expertos en el diseño de consultas el modelo booleano alcanza su máximo potencial.

El **modelo vectorial** (Salton y McGill, 1983) resuelve la limitación de los pesos binarios del modelo booleano introduciendo un grado de similitud entre cada documento de la colección. De esta forma este modelo puede recuperar documentos en los que no hay una coincidencia total con los términos de búsqueda. La relación entre término y documento expresa el grado en que determinado término describe al documento. Los pesos de los términos intervienen en el cómputo del grado de similitud entre cada documento y la pregunta del usuario. Se considera un modelo popular por su alto rendimiento en la recuperación (Türker, 2004).

La mayoría de los motores de búsqueda utilizan una variación del modelo booleano y el vectorial para la ordenación de resultados (Baeza-Yates y Ribeiro-Neto, 1999).

El **modelo probabilístico** (Bookstein, 1983) se basa en la estadística como modo de resolución del problema IR (Moens, 2000). Calcula la probabilidad de relevancia de los documentos para una consulta con el fin de encontrar el conjunto ideal de documentos pertinentes. El valor asignado a cada pareja <*término, documento*> indica la probabilidad de que el documento sea relevante para ese término.

El **modelo difuso** (John y Mooney, 2001) se basa en la teoría de conjuntos difusos (Zadeh, 1965), donde el grado de pertenencia entre un término y un documento expresa la capacidad del término para describir el contenido del documento. También, ha sido denominado modelo booleano extendido y representa una alternativa al modelo vectorial por la flexibilidad en las ponderaciones de los índices.

Otros modelos de recuperación combinan el contenido de los documentos con su estructura. Las búsquedas de información, además de las palabras clave, incluyen el estilo o la estructura asociados a ellas en los documentos.

Una alternativa diferente en IR consiste en los modelos para la navegación. En este caso, los usuarios no realizan consultas. El modo de encontrar la información se basa en la exploración de documentos navegando por enlaces o referencias entre ellos.

Como se ha comentado anteriormente, todos estos modelos, y principalmente el vectorial, ofrecen buenos resultados en términos de exhaustividad y precisión. Sin embargo, como se puede apreciar por las descripciones de los mismos, estos sistemas no dan la necesaria importancia a las características que muestra el entorno web. Este entorno se caracteriza por estar compuesto por miles de millones de documentos muy heterogéneos (tanto por tema, como por audiencia y formato), de naturaleza dinámica, con alto nivel de obsolescencia, y con interacciones significativas entre documentos (mediante hiperenlaces y arquitecturas web). Los modelos tradicionales al no contemplar estos factores dan resultados deficientes. Es por ello que adquieren importancia los sistemas que añaden a estos sistemas de recuperación las características de la web en el posicionamiento de resultados.

2.1.4 Relevancia documental en la Web

La relevancia documental comienza con los trabajos de Cole y Eales (1917). La necesidad de caracterizar la producción científica ha desembocado en multitud de estudios bibliométricos y cienciométricos. Estos análisis se basan en indicadores extraídos de los documentos que permiten establecer relaciones entre documentos para así caracterizar determinada disciplina, grupo (investigadores, institución, país, etc.) o colección documental. Destacan entre los indicadores más representativos el análisis de citas y referencias.

Aunque inicialmente menos conocido que el análisis de citas y referencias, los estudios lingüísticos y estadísticos de los términos de los documentos para caracterizar determinado corpus tienen también una larga tradición. Así lo atestiguan los estudios de coaparición de palabras de Callon (1993, 1995); de Seglen (1996), que recoge análisis estadístico de las palabras de las páginas, temas e ilustraciones; de Gilyarevsky (1997) que se centra en el estudio lingüístico de los términos del título; Losee, (1996), que estudia las estadísticas de ventanas de texto; y Morato (1999) que confirmó que las estructuras documentales, el dominio temático y el tipo de audiencia a la que un documento va dirigido tienen efectos sobre la calidad de la representación del dominio que estos describen.

La relación entre la calidad documental y los indicadores arriba mencionados, ha sido estudiada en diferentes investigaciones. Así Lawani (1986) confirmó la correlación positiva entre la calidad de la investigación y los indicadores bibliométricos, como el número de autores por publicación, la productividad del país, o el número de citas en un periodo de cinco años. Los indicadores recogidos por los autores han sido correlacionados con el indicador que tradicionalmente se ha considerado aval de calidad, es decir, la opinión de los expertos. Si bien estos resultados no pueden ser extrapolados sin más de un área a otra. Así, Smart (1983) encontró que la correlación entre calidad y citas era menor en el campo de la educación que en otras áreas.

El número de citas es evidentemente una medida de la visibilidad. Sin embargo, la relación entre calidad y citas muestra conclusiones contrapuestas. Por ejemplo, mientras Cole y Cole (1971), Clark (1957) y otros trabajos recogidos por Egghe y Rousseau (1990) defienden que las citas evidencian calidad documental, Pontigo y Lancaster (1986) descartan esta hipótesis.

A pesar de las controversias, estos indicadores se siguen utilizando, e incluso se han extrapolado o adaptado con objeto de aplicarlos a documentación no científica. Es el caso del análisis de citas, que tiene su versión para documentación web en el análisis de enlaces (Chau, 2003) y se traduce en una forma de evaluar la visibilidad y el impacto (Martin e Irvine, 1983). Se pueden establecer más semejanzas con el resto de indicadores y sus posibles aplicaciones a los recursos web e indicar que comparten principios como la obsolescencia de la información. No obstante, la heterogeneidad de temáticas, tipologías y usuarios, la diferente dispersión literaria (Bradford, 1948), unido al inmenso número de documentos dinámicos en permanente competencia, hace inviable la aplicación directa de los métodos tradicionales de relevancia documental a la Web.

La relevancia documental en la Web determinada por el posicionamiento web depende directamente de las características asociadas a los documentos que son consideradas por los motores de búsqueda (Marckini, 2001). Estas características se pueden identificar por el orden de los documentos (páginas web o recursos web) ofrecidos como respuestas por los buscadores ante consultas de sus usuarios.

Son muchas las cuestiones principales a la hora de entender la relevancia documental en la Web, sobre todo si se tiene en cuenta la presencia de multitud de documentos poco fiables y no estructurados donde el significado de la información no siempre es evidente (Machill et al., 2003). Ejemplos destacados de estas cuestiones son:

- ¿Cómo ordenan las páginas web los buscadores?
- ¿Cómo se puede mejorar el posicionamiento de una página web?
- ¿Qué herramientas nos pueden ayudar a mejorar el posicionamiento web?
- ¿Cómo actúan y que contemplan estas herramientas de mejora del posicionamiento web?

Desde esta perspectiva, el estudio de la relevancia web se traduce en la comprensión de los algoritmos de posicionamiento de los motores de búsqueda web.

2.1.5 Búsqueda de información Web

Los directorios y motores de búsqueda son las dos plataformas principales que se utilizan para buscar y recuperar información en la Web. Se diferencian principalmente por su forma de indexar y recuperar los documentos. En los directorios son editores humanos los que asignan categorías temáticas a los documentos mientras que los motores de búsqueda

los indexan de forma automática. Es decir, en un directorio la indización es por asignación intelectual de un término, presente o no en el recurso, mientras que en un motor la indización es por extracción de los términos del recurso. A su vez, la recuperación de información en los directorios se realiza apoyándose en un modelo de navegación que permite moverse por recursos de Internet clasificados en árboles jerárquicos de categorías (Babiak, 1999). Los motores de búsqueda siguen un esquema de recuperación que puede tratar un mayor número de documentos para encontrar aquellos que son pertinentes (Chang et al, 2001), en comparación con los directorios (Ferber, 2003).

Además en los motores de búsqueda, los documentos se recuperan de acuerdo a los términos de la consulta. Para ello, un algoritmo de clasificación determina qué documentos son pertinentes, presentando los que considera más relevantes en puestos preferentes.

Tanto los directorios como los motores de búsqueda tradicionales llegaron a la Web con muy poca diferencia temporal. El primer directorio, Yahoo, tuvo su origen en 1994 (Sullivan, 2003) mientras que Altavista, el motor de búsqueda convencional más antiguo, apareció en el mercado en el año 1995. Sin embargo, los motores de búsqueda han alcanzado mayor popularidad que los directorios convirtiéndose en los sitios web más visitados en Internet.

A pesar de las distinciones hechas entre directorios y motores, realmente no tienen un funcionamiento independiente en la práctica. Los buscadores web suelen apoyarse en la información estructurada de los directorios y en los resultados que provienen de otros motores de búsqueda para recuperar información. Por ejemplo, Google utiliza el directorio colaborativo DMOZ para establecer una estructura clara y ordenar sitios web por importancia (Türker, 2004). Hasta hace poco tiempo el directorio Yahoo redireccionaba al motor Yahoo!Search cuando no encontraba resultados ante una consulta.

Además de existir interacciones entre los resultados de motores y directorios, también existen interacciones entre motores. Algo particularmente evidente en los denominados metabuscadores, buscadores que combinan los resultados de diferentes buscadores. En la siguiente figura se muestran este tipo de interacciones entre motores y dependencias entre

los principales recuperadores de información. En esta figura, las líneas rojas muestran que existe una relación entre los primeros resultados posicionados en el buscador, las líneas amarillas representan resultados comunes en enlaces patrocinados.

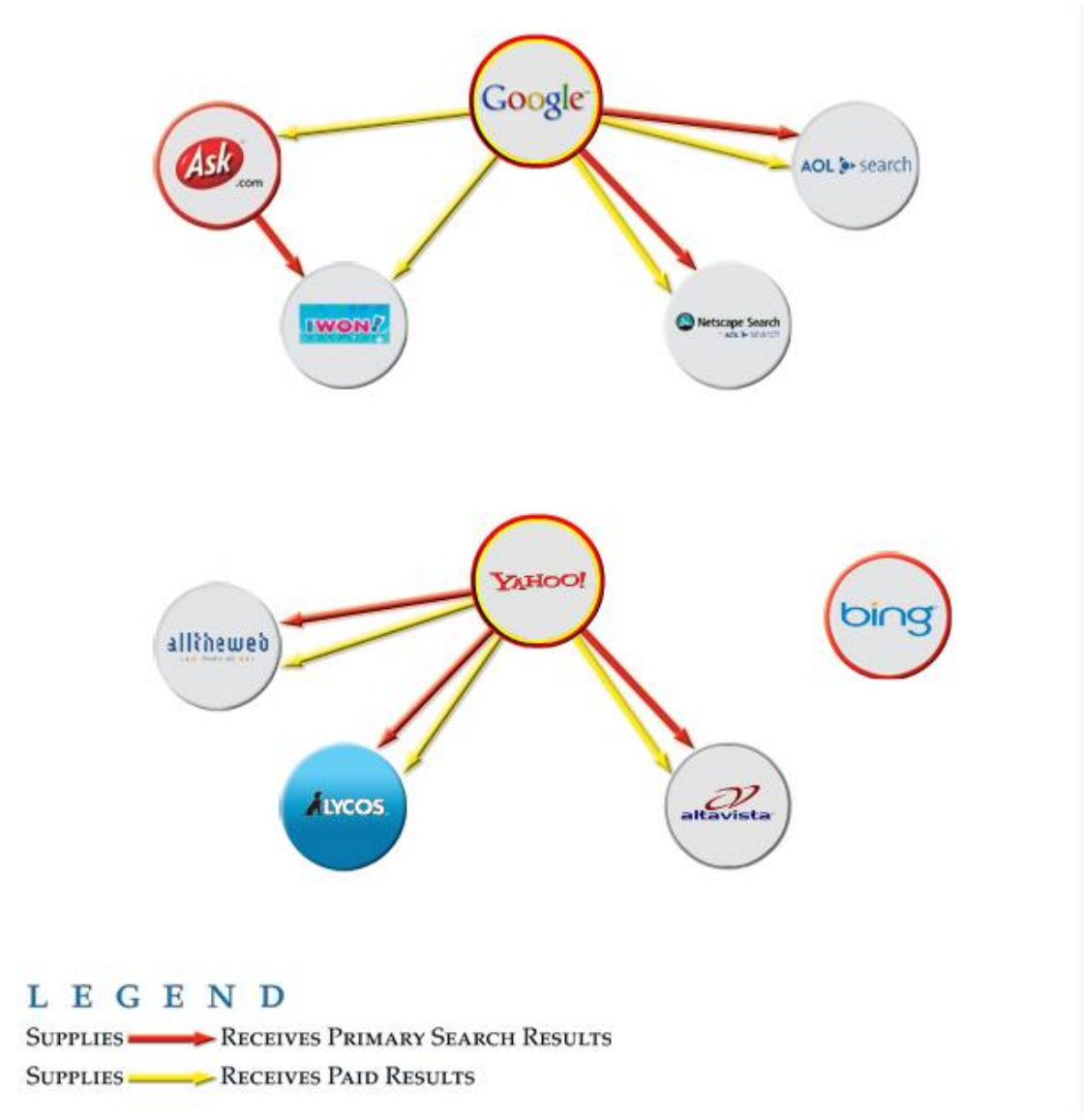


Figura II-2: Relaciones entre motores de búsqueda web tomado de Clay (2009)

2.1.5.1 Arquitectura de los buscadores web tradicionales

La arquitectura de un motor de búsqueda se basa en tres módulos principales: un módulo de control de rastreado (también denominado araña o crawler), un módulo de indexación, un motor de consulta y un modulo de clasificación. La siguiente figura muestra la arquitectura típica de un motor de búsqueda (Arasu et al., 2001). Si bien pueden existir ligeras variaciones en la organización y complejidad del motor.

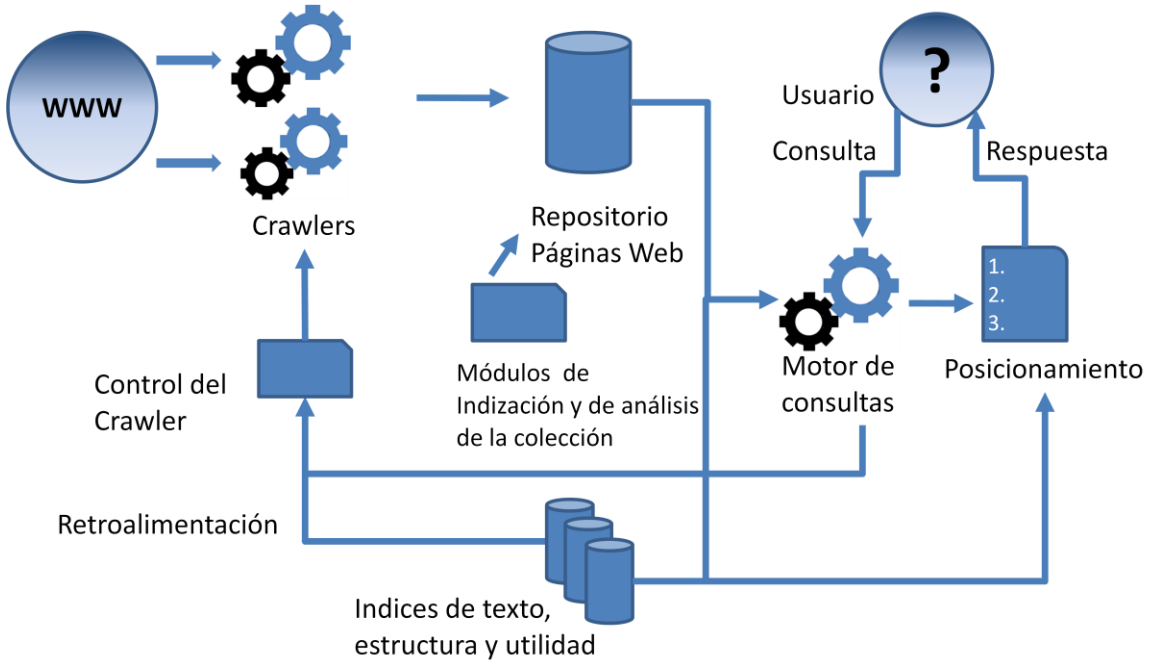


Figura II-3: Arquitectura de un motor de búsqueda tomado de Arasu et al. (2001)

El rastreador web o crawler se inicia en un determinado conjunto de direcciones URL y sigue los enlaces para acceder a otras páginas (Bradmanet et al., 2000). Las URLs que aparecen en los documentos recuperados son analizadas por el "módulo de control del crawler", que determina los vínculos que se visitarán por el rastreador. Los documentos obtenidos se almacenan en el repositorio de páginas web. El "módulo de indexación" extrae todos los términos de los documentos y les asigna las direcciones URL donde se han encontrado. El "módulo motor de consultas" procesa los resultados de búsqueda en función de los distintos índices del repositorio, y finalmente son ordenados por relevancia por el "módulo de clasificación".

2.1.5.2 Estrategias avanzadas de búsqueda web

A medida que la Web se expande, con sitios escritos en lenguaje natural, la búsqueda de información específica se convierte en un reto cada vez más difícil (Choi et al., 2003).

Los motores de búsqueda de propósito general no siempre generan resultados con un grado de pertinencia adecuado. Los motores de búsqueda específicos se centran en un número menor de temas. La simplificación del problema de IR les da ventaja, en su campo, sobre los grandes buscadores generalistas (Giles et al., 2003). Los buscadores

temáticos se basan en la información recogida por herramientas de filtrado que miden la relevancia de los documentos (Chang et al., 2001).

Los buscadores destinados a la localización de productos, servicios, precios y proveedores en Internet son motores especializados orientados al comercio web. Los grades motores de búsqueda tienen asociados algún buscador que ofrece este tipo de servicios. Es el caso de Froogle, vinculado a Google, que permite la búsqueda de productos mediante términos de búsqueda (Paulson, 2003), y también el de Yahoo! Shopping.

La metabúsqueda es otro enfoque diferente dentro de IR. Los metabuscadores combinan los resultados de varios motores de búsqueda para responder a una demanda de información. Dicho de otra manera, el usuario puede plantear una consulta en diversos motores de búsqueda a través de una única interfaz. (Schwartz, 2001). Cada metabuscador tiene su propio algoritmo de combinación de resultados. Casi todos los metabuscadores tienen un tiempo máximo de espera de respuesta para los distintos motores. Su éxito se debe a que los motores de búsqueda no indexan la totalidad del contenido de los sitios web (Baeza-Yates y Ribeiro-Neto, 1999), y además el contenido indexado por cada buscador web suele ser diferente. Por tanto, la combinación de resultados debe ser una mejor representación del contenido real. Estos metabuscadores tuvieron un momento de auge en los primeros años de la presente década, para luego declinar a favor de motores únicos. Actualmente, experimentan un nuevo empuje debido a las búsquedas en repositorios multimedia de la Web 2.0.

MetaCrawler es un ejemplo de este tipo de recuperadores de información. Fue creado en 1995 en la Universidad de Washington y posteriormente adquirido por InfoSpace en el año 1997. Otros metabuscadores con técnicas avanzadas de organización de los resultados mediante agrupación son KartOO y Mooter (Dvorak, 2004), incluso algunos permiten su instalación local, como Copernic .

Una de las posibles mejoras a los motores de recuperación lo constituyen los sistemas de pregunta-respuesta. Estos sistemas permiten interrogar en lenguaje natural al sistema, evitando que el usuario tenga que aprender un lenguaje de consulta (Chang et al., 2001). A diferencia de los buscadores tradicionales estos motores responden con la respuesta a la consulta (Zahdeh, 2003), en vez de con un conjunto de documentos que probablemente

contienen la respuesta en su contenido (Lin et al., 2003) . Las respuestas se deducen a partir de una base de conocimiento (Kwok et al., 2001) confeccionada por múltiples fuentes (Katz et al., 2002). La extracción de la información se realiza para identificar entidades relevantes en los textos (Cowie y Wilks, 2000). Esta tarea está frecuentemente asociada a los sistemas pregunta-respuesta pues permite la identificación de unidades de información que normalmente forman parte de las consultas.

En una línea similar, en cuanto a razonamiento sobre el contenido de los sitios web (Bozsak et al., 2002), pero mucho más ambiciosa, se sitúa el concepto de Web Semántica (Berners-Lee et al., 2001), (Shah et al., 2002). La idea subyacente es mejorar los servicios que provee la Web mediante una mayor formalización de la semántica contenida en los recursos que la integran.

La Web Semántica utiliza fuentes adicionales para incorporar semántica mediante ontologías (Stuckenschmidt, 2002), (Doan, 2002) y vocabularios de metadatos.

Cualquiera de los enfoques expuestos se pueden combinar con respuestas humanas para paliar dificultades en la automatización de IR (Sherman, 2003).

Muchos de los avances en el campo de la búsqueda web están orientados a mejorar la usabilidad. El objetivo perseguido es la recuperación de recursos de calidad por parte de los usuarios, sin obligarlos a aprender un lenguaje de consulta (Eastman y Jansen, 2003) ni a hacer demandas de información correctas (Kline, 2002). En este sentido se han propuesto métodos más intuitivos para la formulación de las consultas (Machill, 2003). Estos métodos agrupan los resultados de las búsquedas por palabras, frases y categorías (Bergstein, 2004), pudiendo incorporar también información gráfica en forma de mapas contextuales (Machill et al., 2003). La organización conseguida, en comparación con las típicas listas de resultados que devuelven la mayoría de los buscadores (Kobayashi y Takeda, 2000), hace efectivas las consultas con pocas palabras, simplificándose así la construcción de consultas.

2.1.5.3 Evaluación de los buscadores web

La efectividad de los algoritmos de búsqueda es un tema de interés para los usuarios (Hawking et al., 1999). La eficacia en la recuperación de un buscador web se mide en función de la pertinencia de los resultados que proporciona (Robertson, 2000), aunque es un concepto difícil de definir (Ingwersen, 2000). El proceso de evaluación concluye con dos medidas, la *precisión* (grado de pertinencia de los documentos web recuperados) y el *recall* (cantidad de información pertinente a disposición del usuario).

2.1.5.4 Limitaciones en la accesibilidad a la documentación web

Los buscadores web sólo pueden acceder a una parte de la Web debido a múltiples factores (Baeza-Yates, 2003). En algunos casos, los robots de indexación de los motores de búsqueda ignoran la existencia de algunos documentos. Hay que tener en cuenta que los crawlers localizan los documentos mediante enlaces o por las URL que son dadas de alta en el site del buscador, una página sin enlaces y sin dar de alta puede ser difícil de localizar. Otras veces se accede al recurso pero el tratamiento de su contenido requiere una herramienta especializada (Ludwig, 2003), también puede suceder que el documento esté protegido. Ejemplos de estos documentos son los que solo son accesibles mediante contraseña, los que requieren rellenar un formulario (p.e. OPACs), las imágenes con texto o los PDF con el contenido codificado. El conjunto de documentos inaccesibles para los buscadores web se denomina Web invisible o profunda.

La mayor parte de la información inaccesible se encuentra en bases de datos profesionales (McGuigan, 2003), y por tanto, suele superar en calidad a los contenidos visibles (Machill et al., 2003).

2.1.6 Algoritmos de posicionamiento

Los algoritmos de posicionamiento se basan en fórmulas matemáticas que calculan valores numéricos para la ordenación de los resultados ante una determinada consulta. Estos algoritmos pueden utilizar cientos de factores como: la importancia de una página; las veces que se cita; la ocurrencia de una palabra; dónde está situada esa palabra en el contexto de la página, etc. En la actualidad suponen un elemento imprescindible para todos los sistemas de recuperación, y en especial para los buscadores web (Kobayashi y Takeda, 2000).

Los algoritmos de posicionamiento de los motores de búsqueda deben tener presentes algunos aspectos. Quizás uno de los más críticos para su estudio es que los algoritmos de posicionamiento, que difieren de un buscador a otro, se basan en secretos comerciales, (Roebuck, 2000). Este secreto comercial es la base de su negocio, y es clave para obtener algoritmos de búsqueda y posicionamiento mejores que los de sus competidores. Además, modifican periódicamente sus algoritmos para defenderse de posicionamientos fraudulentos.

Cada motor de búsqueda elabora su propio algoritmo de posicionamiento. Algunos de los algoritmos más populares son el PageRank de Google, el WebRank de Yahoo Search o el del buscador Bing de MSN.

Como indica R. Baeza-Yates (1999), el dinamismo del posicionamiento web, donde compiten entre sí documentos y otros recursos web por estar mejor posicionados, y por otro lado las modificaciones en los algoritmos de posicionamiento de los buscadores web, hace que no se perciba el posicionamiento como una ciencia exacta sino como un conjunto de heurísticas de eficacia variable y condicionada en el tiempo. Estrategias validas en un determinado momento quedaran desfasadas conforme evolucionen los algoritmos. Esta competición ha sido descrita como un ejemplo de "Efecto de la Reina Roja" (Morato et al., 2005). De aquí la necesidad de metodologías automáticas adaptables a estas evoluciones.

2.1.6.1 PageRank

Su mayor innovación fue incorporar la relevancia a los enlaces entrantes que apuntaban a las páginas, en lugar de calcular la relevancia y la pertinencia de un resultado para una consulta sólo en función de la densidad de palabras clave. Para determinar el PageRank (PR), Google analiza el número de enlaces que provienen de otras páginas web (Lawrence y Brin, 1998). Su lógica es la siguiente (Craswell et al., 2001), si una página web enlaza con otra página, es que la está recomendando. Y si la recomienda, es muy probable que sea importante en el ámbito del tema que trata la primera página web (Richardson y Domingos, 2004). Por este motivo a las páginas que reciben muchos enlaces se les considera de mayor calidad (Chang et al., 2001) y (Walker, 2002).

En el cálculo del PageRank (Chen et al., 2002) se tiene en cuenta el PageRank de las páginas de las que provienen los enlaces. Es decir, los enlaces no se valoran por igual si

no que dependen de la "calidad" de la página que lo emite. Y esta "calidad" viene dada a su vez por el número de páginas que apuntan a esa página (además de otros factores que luego son incorporados).

A pesar del éxito del PageRank se ha demostrado que es una medida sesgada que perjudica a las páginas de reciente incorporación a la Web (Baeza-Yates et al., 2002), por lo que se han propuesto en el mismo trabajo modificaciones del PageRank para solventar este problema. A su vez, en (Pretto, 2002) se aconseja incluir en el PageRank la percepción de los usuarios sobre la calidad de las páginas web.

Los creadores del algoritmo de Google (Brin y Page, 1998) aseguran que existen más de 100 factores o variables que son considerados en el algoritmo de posicionamiento. Algunos de estos factores son:

- Total enlaces entrantes, es decir enlaces de otras páginas web.
- Enlaces entrantes de webs con PR4 o mayores. Cuanto mayor es el PR de la web que nos enlaza, mejor.
- Palabra clave en el texto del enlace, la manipulación de este factor provoca el fenómeno *bombing*.
- Número de enlaces externos que enlazan a las páginas que nos enlazan.
- Posición del enlace en la página que enlaza. Cuanto más al inicio en el código HTML mayor importancia.
- Densidad de palabras clave en la página que enlaza.
- Título HTML en la página que enlaza.
- Enlace de sitio de "experto" (webs con listas de recursos de una misma temática). Estos enlaces son relevantes, proceden de fuentes reputadas sobre la misma temática.
- Temática de la página que enlaza. Mejor si es de la misma temática que la nuestra.
- Estar incluido en una categoría de DMOZ[1].

[1] DMOZ (http://www.dmoz.org) es el directorio más grande de la Web y es mantenido por editores voluntarios. Fue adquirido por Netscape (AOL) en 1998. La información del directorio puede ser utilizada por cualquier persona a través de un acuerdo de licencia abierta (Ferber, 2003).

2.1.6.2 HITS

El algoritmo Hypertext Induced Topic Search (HITS) fue propuesto por Kleinberg (1999). Se trata de un método de clasificación basado en la conectividad que, a diferencia del PageRank, es dependiente de la consulta del usuario.

A partir de resultados de búsqueda iniciales incluye nuevos resultados que enlazan o son enlazados por los primeros. Identifica en el grafo de páginas web los nodos *autoridad* (los que reciben más enlaces) y los "*hub*" (los que apuntan a muchas páginas pertinentes) mediante un sistema de pesos. Las páginas *autoridad* son las que tienen más posibilidades de ser relevantes para la consulta realizada (Henzinger, 2000), (Diligent et al., 2002). Los "*hub"* no son necesariamente las propias autoridades, pero al menos permiten acceder a las autoridades. Esto crea una doble vía de retroalimentación positiva entre autoridades y los "*hub"* (Baeza-Yates et al., 2002). En (Schimkat, et al., 2002) se describe el proceso de búsqueda de los sitios Web elitistas en cuanto a su pertinencia para el tema de consulta. Se trata de utilizar sólo las páginas autoridad en lugar de todos los documentos considerados como pertinentes.

2.1.6.3 WebRank

Los webmasters a través de los foros aseguran que para el algoritmo de Yahoo Search, WebRank, los factores que más condicionan el valor para el ranking son:

- Enlaces entrantes.
- Estar dado de alta en directorios de Yahoo! y DMOZ. Aunque en estos momentos el directorio de Yahoo! está desactualizado y no con todas sus funcionalidades habilitadas al menos en la versión española.
- Igualdad entre el término de consulta y el término encontrado en el texto, teniendo en cuenta que no se usan listas de palabras vacías.
- Yahoo proporciona relevancia al título de la página.

2.1.6.4 TrustRank

El TrustRank (López, 2009) indica la credibilidad de los sitios web. Para una página web su prestigio se basa en la calidad de su contenido y en el TrustRank de las páginas que la enlazan. No obstante, una página web origen de enlaces puede indicar al motor de búsqueda que no transmita a una página vinculada por ella su fiabilidad.

Los sitios con mejor TrustRank son organismos oficiales, universidades, medios de comunicación contrastados, etc.

2.1.6.5 Técnicas de ordenación basadas en el contenido documental

Entre las técnicas más utilizadas figuran las de filtrado terminológico (Frakes y Baeza, 1992) como, por ejemplo, las basadas en la frecuencia en la colección, la frecuencia en el documento, IDF (*Inverted Document Frequency*) y las frecuencias de cadenas (n-grams).

- **Frecuencia en la colección**

  Los trabajos de Zipf (Cleveland, 1990) constituyen las primeras investigaciones sobre filtrado terminológico a partir de las frecuencias de los términos en la colección. Zipf defendía el principio del mínimo esfuerzo para comunicarnos (Zipf, 1949), es decir, nos comunicamos con el menor número de palabras que sea capaz de preservar el significado del mensaje. Según este principio las palabras de uso más frecuente deben ser cortas.

  Continuando los trabajos de Zipf, Goffman (Chaumier y Dejean, 1990) considera que las palabras de mayor frecuencia se corresponden con palabras funcionales de escaso valor semántico, las de menor frecuencia muestran el estilo del autor y las de frecuencia intermedia son las palabras que aportan la semántica al documento. Otro estudio (Jones, 1992), indicó que las palabras de baja frecuencia contienen mayor carga semántica mientras que las palabras no vacías de frecuencias superiores tienen relación con la temática del documento. Posteriormente, Spyns (Spyns et al., 2004) determinó que las palabras de baja frecuencia eran útiles para determinar el vocabulario técnico del dominio. En (Spyns y Reinberger, 2005) se realza la utilidad de términos con baja frecuencia trabajando con tripletas de términos. Aunque Price (Price y Thewall, 2005) añade que las palabras de muy baja frecuencia no siempre son descartables.

  A la vista de las diferentes interpretaciones se pone de manifiesto la falta de precisión del método. Para paliar esta deficiencia se ha combinado su uso con tesauros, *algoritmos de stemming*[2], medidas de coocurrencia de términos y ponderaciones de los términos según su ubicación en la estructura del documento.

[2] algoritmos que reducen las palabras a su raíz eliminando los afijos y agrupando todas aquellas que comparten la misma raíz (Lovins, 1968)

- **Inverted Document Frequency (IDF)**

  En 1972 Spark-Jones definió el método IDF para asignar pesos a los descriptores (Moens, 2000). Este método mide la frecuencia relativa de un término respecto al conjunto de documentos de la colección (Cleveland, 1990). Tal y como está definido cuanto mayor sea el número de documentos en los que aparece un término menor será su poder de discriminación (Salton, 1989), (Schultz, 1968). Desde este punto de vista, los términos seleccionados por este método son adecuados para indizar y recuperar documentos pero en general no está garantizada su pertenencia al vocabulario controlado de un dominio.

- **Frecuencia en el Documento (TF)**

  La frecuencia en el documento es el número de ocurrencias de un término en el documento (Glöggler, 2003). En combinación con IDF obtiene mejores resultados, compensando las frecuencias altas del término y la longitud del documento (Harman, 1994), (Moldovan y Surdeanu, 2003).

  Se debe tener en cuenta que estas técnicas son más o menos adecuadas en función de la finalidad de la identificación de términos.

  Por ejemplo:

  - recuperación de información (Salton, 1989)
  - indización automática (Hodges et al., 1996)
  - extracción de resúmenes (Paice, 1990); (Luhn, 1958)
  - similitud entre documentos (Hatzivassiloglou et al., 1999)

- **Frecuencias de cadenas (n-grams)**

  Este método estadístico (Cohe, 1995) consiste en representar un documento por las secuencias de caracteres consecutivos de longitud n (natural prefijado) que aparecen en el documento junto con su frecuencia de aparición.

  A estas cadenas se les denomina n-grams. Una de las técnicas de procesamiento de n-grams consiste en comparar los n-grams extraídos con los de un background construido para tener elementos comparativos. Este background se crea con un conjunto de documentos pertenecientes a un dominio lo más disjunto posible del dominio que se pretende modelar. En función del background variará el grado de especificidad de los descriptores obtenidos (Velasco et al., 1997).Una vez

comparados y puntuados los n-grams se asignan pesos a los caracteres basándose en los n-grams a los que pertenecen. A partir de estos pesos se seleccionan los descriptores (palabras o frases) atendiendo a los caracteres que los forman.

2.1.7 Factores relevantes de Posicionamiento

Los motores de búsqueda comparten similitudes en cuanto a los factores relevantes de posicionamiento. La mayoría de las características que influyen en el posicionamiento son tenidas en cuenta por los buscadores, aunque cada buscador pondera las características de forma diferente e incluso modifican cada cierto tiempo los pesos para evitar técnicas de cálculo exacto del posicionamiento.

Los factores también conocidos como atributos, características, o variables SEO que afectan al posicionamiento admiten diferentes clasificaciones (Sullivan, 2007; Morato et al., 2009). Los factores relacionados con las características propias de la página web se dice que son factores internos (On Page) del contenido de la página y los relacionados con cualidades externas y del entorno se denominan factores externos (Off Page) (Sullivan, 2007). Desde el punto de vista de la evaluación que realiza un buscador se dividen en factores directos, los que evalúa el motor de búsqueda, y factores indirectos, aquellos que no se evalúan por el motor, pero afectan a la evaluación de los factores directos (Morato et al., 2009). En cualquier caso, tal como indica Fuhr (2000), la recuperación de información además del contenido de los documentos, debe tener en cuenta la estructura lógica, la disposición, y los atributos externos a los documentos, así como la participación de los usuarios

2.1.7.1 Factores directos

Se entiende por factores directos a aquellos atributos que evalúa el motor de búsqueda en el cálculo de la relevancia web de un documento, como la popularidad de una página, los perfiles de usuarios, las características del sitio web, etc. (Morato et al., 2009). A continuación se definen los principales factores directos encontrados.

2.1.7.1.1 Popularidad de la página

Los motores de búsqueda asumen que la popularidad de una página es un indicio de su calidad. La popularidad de una página web mide la consideración que tienen los usuarios u otras páginas web sobre una página.

Una página o un sitio web tiene más popularidad, desde el punto de vista de los usuarios, cuanto mayor número de visitas recibe. Los buscadores consideran que una web muy visitada contiene información de interés general, y por tanto es muy probable que interese al usuario que realiza una consulta. Hay diversas maneras de medir este tipo de popularidad (Machill et al., 2002): número de visitas absolutas, número de visitas relativas[3] , números de clics, tasa de clics (*clickthrough*)[4], alcance (*reach*)[5], páginas visitadas[6] (*page views*), tráfico del sitio[7].

La popularidad adquirida por el reconocimiento de otras páginas web se valora por el número de enlaces externos que recibe (Henzinger, 2001) y/o la calidad de las páginas de origen de esos enlaces. El PageRank (Page et al., 1998) que utiliza Google y los HITS (Kleinberg, 1998) son los ejemplos más representativos de formas de medir este factor (Richardson y Domingos, 2002), (Wensi et al., 2002), (Agosti y Melucci, 2000). Otras variantes de estos algoritmos se recogen en (Marendy, 2001), (Lawrence y Giles, 1999).

Los factores de popularidad son de gran interés para la mayoría de los buscadores no sólo por ser indicadores de autoridad documental sino también por su relativa dificultad a la hora de intentar manipularlos.

Los webmasters han restado eficacia a estos factores al aumentar la contaminación mediante la utilización de políticas de intercambios de enlaces y granjas de enlaces (*link farms*) (Morato et al., 2005). Las granjas de enlaces o *link farms* son acuerdos entre varios sitios web para incluir las páginas web en páginas de enlaces que las apunten (Henzinger, 2000a).

Como contrapartida, los motores de búsqueda se protegen de las alteraciones fraudulentas en los factores de popularidad mediante estrategias de penalizaciones. Por ejemplo, en el trabajo (Moreno, 2005) se evidenciaba que Google penalizaba aquellas páginas cuyo número de enlaces externos no estaba avalado por el tráfico de visitas que recibía.

[3] Número de visita recibidas por una página comparadas con las de páginas de su competencia
[4] Clics en los que el usuario recibe la página a la que desea acceder
[5] Número de usuarios (direcciones IP) que visitan un sitio en un determinado día.
[6] Cantidad de páginas visitadas por las urls diferentes que visitan un sitio. En distintos días la misma url se cuenta como diferente.
[7] Es una combinación del alcance y las páginas visitadas. Determina el número de usuarios que se interesaron por un sitio web durante un periodo concreto.

2.1.7.1.2 Perfiles de usuario

Las características de los usuarios deben ser tenidas en cuenta en el diseño de páginas web si se desea conseguir una buena optimización. Los motores de búsqueda pueden conocer el perfil del usuario mediante códigos de seguimiento en la página o por medio de *logs* de acceso al servidor (Sullivan, 2003a).

Los buscadores asignan más relevancia a las páginas cuya temática se corresponde con la preferencia del usuario (Jeh y Widom, 2003). También tienen en cuenta en la ordenación documental el idioma del usuario, intuido a partir de la propia consulta, la localización geográfica y el idioma instalado por defecto en su máquina. Otra forma de presentar la información para facilitar el ajuste a un perfil de usuario es mediante módulos de contenido (Rossi et al., 2001). Yahoo Search procede de esta manera filtrando el contenido de los módulos de acuerdo a la información personalizada de los usuarios.

En el caso de consultas que son demandas de servicios la proximidad geográfica es considerada por los motores de búsqueda como un aporte de relevancia (Sterling, 2004).

Las cuestiones de privacidad han limitado el uso de estas técnicas, ya que los usuarios temen que su datos puedan llegar a terceros (Thompson, 2003). Ejemplos como el de Gmail (Sullivan, 2004), cuyos correos privados son explorados por Google con el propósito de incluir anuncios relevantes, intranquilizan a los usuarios.

2.1.7.1.3 Intereses económicos

Un sistema de promoción muy utilizado por los sitios de venta de productos en línea consiste en el pago por clic (*Pay-per-click*). Los propietarios del recurso web a promocionar pagan una cantidad económica cada vez que se ejecuta un clic de acceso a su sitio web. El posicionamiento dependerá del precio ofrecido por clic.

Otra forma de intentar obtener un buen posicionamiento mediante abono directo de cantidades económicas es el pago por inclusión (*pay-for-inclusion*). Se paga en este caso por estar incluido en un conjunto de posiciones relevantes ante determinadas consultas. El orden exacto dentro de este conjunto dependerá de la relevancia de los contenidos.

2.1.7.1.4 Características del site

Las características del site pueden repercutir en la fiabilidad del mismo. Los dominios pertenecientes a organizaciones estatales como *.gov*, *.edu*, *.org* o los de ámbito global en la red como *.com* o *.net* llevan asociados mayor credibilidad. También el tamaño del site puede repercutir en su prestigio.

2.1.7.1.5 Novedad del contenido

Las páginas de creación reciente y aquellas que se actualizan con frecuencia adquieren ventaja en el posicionamiento al considerarse que su información es más actual (Thompson, 2002). Este factor adquiere aún mayor importancia en ámbitos en los que existe gran dinamismo en la información, como por ejemplo en los noticiarios y la meteorología. No obstante, los sitios web que sin ser novedosos permanecen fieles a una misma temática, también son considerados relevantes por los motores de búsqueda.

2.1.7.1.6 Formato y código de página

Con respecto al formato y codificación de la página, existen varios factores que afectan. Esos factores se exponen a continuación.

- Las **etiquetas Meta** (META Tags) incluyen las palabras clave de las páginas y describen su contenido (Thompson, 2002). Los buscadores como Google apenas las tienen en cuenta (Sullivan, 2002) posiblemente por ser objetivos de fácil manipulación, los índices tampoco las suelen consultar adquiriendo esta información de los formularios utilizados al registrar las páginas. A pesar de ello, las estrategias SEO recomiendan crearlas de forma adecuada, destacar que buscadores como Infoseek o Inktomi si las consideran relevantes.

- Las características del **texto de los enlaces** tales como su número de palabras son tenidas en cuenta la hora de posicionar las páginas.

- El **formato de texto** (tamaño, negrita, cursiva…) y los encabezados (*headings*) de los apartados y subapartados influyen en el cálculo de la relevancia de los resultados (Thompson, 2002) y (Ferber, 2003). Los elementos de encabezado que dan estructura a los documentos son seis (H_n, con n=1..6). La codificación de la página se tiene en cuenta en el posicionamiento a partir de la aparición de las siguientes etiquetas:

  - Javascript: indica que el código se ha implementado en JavaScript (lenguaje interpretado de sintaxis semejante a Java orientado a las páginas web).

- Noscript: los browsers que interpreten JavaScript e incorporen esta etiqueta ignorarán el código HTML encerrado por la misma. Sin embargo será mostrado como alternativa por otros browsers que no entiendan "JavaScript".
- Noembed: permite dotar a una página web de contenido alternativo a los objetos "embebidos", o incrustados en la página web.
- Noframe: advierte a los buscadores que no indexan marcos (frames) que consideren su texto en la indización.
- Frameset: etiqueta clave para la creación de los frames de las páginas.
- Flash: indica inclusiones de animaciones flash en la página web.

El número de estas etiquetas en el documento y el porcentaje de código con respecto al número de palabras visibles del documento también son valorados en el posicionamiento.

2.1.7.1.7 Semejanza entre términos de la consulta y términos en la página web

Este factor de coincidencia de términos es el que mejor evalúa la correspondencia entre el contenido de un recurso web y la información demandada por el usuario. La coincidencia entre términos se lleva a cabo en distintos apartados del documento: Título, URL, etiquetas asociadas a las imágenes e hiperenlaces que contiene la página web (etiquetas ALT y TITLE), etiquetas META y en el cuerpo del documento (BODY). Las características que se tienen en cuenta en cada uno de estos apartados son:

- Título:
  - Número de palabras del título.
  - Aparición de los términos de búsqueda (Ferber, 2003)
  - Porcentaje de los términos de búsqueda.
- URL:
  - Aparición de los términos de búsqueda en la URL del documento.
  - Aparición de los términos en las URL´s de los enlaces del documento.
  - Número de niveles de la URL (número de apartados que contiene el sitio).
- Etiquetas ALT y TITLE:
  - Número de palabras de la etiqueta.

- Aparición relativa de los términos de búsqueda.

- Etiquetas META:
  - Número de palabras.
  - Porcentaje de términos de búsqueda.
- BODY:
  - Posición del primer término de búsqueda encontrado en el BODY. (Türker, 2004)
  - Aparición de los términos de búsqueda en las primeras cien palabras (Moss, 2001).
  - Porcentaje de aparición de los términos de búsqueda en las primeras cien palabras y en el total.
  - El número de palabras y el número de palabras diferentes del documento.
  - Aparición de los términos de búsqueda en negrita, en mayúsculas o en las cabeceras declaradas de la página web (H_n).
  - Número de palabras en el texto de los enlaces.
  - Porcentaje de aparición de los términos de búsqueda en el texto del enlace.
  - Número de enlaces que salen de la página web.

Como se puede observar, el BODY es el apartado en el que se contempla un mayor número de variables relacionadas con la semejanza de términos de búsqueda. Esto es debido, tanto a la extensión del cuerpo del documento (bastante mayor que el resto de apartados), como al hecho de que se encuentre diversidad de información, como definición de listas, tablas, tipos de escritura, color del texto, etc.

2.1.7.2 Factores indirectos

Los factores indirectos son aquellos que no analiza directamente el buscador, pero condicionan la percepción de los recursos web por parte de los usuarios. Repercuten, por tanto, en los factores directos de popularidad que evalúan los motores de búsqueda.

La imagen que ofrece un recurso web ante los usuarios se debe principalmente a:

- **Estructura de la página web y/o del site**. El contenido debe estar estructurado y organizado. El usuario debe poder intuir el contenido en un primer vistazo, ya que existen unas pautas que los usuarios tienen en cuenta a la hora de buscar información.
- **Credibilidad o fiabilidad** de la página o del site. Se estudia la forma y la relación con la credibilidad: nivel de compromiso, enlaces recibidos y ofrecidos, diseño web.
- **Usabilidad** de la página web mide la facilidad de uso. Las propiedades de la página web que pueden influir en la usabilidad y el manejo de la página son: la facilidad de aprendizaje, la facilidad de recuerdo, la eficiencia del usuario, la tasa de errores y la satisfacción.
- **Accesibilidad** de la página web tiene como objetivo no poner barreras a la audiencia a la que va dirigida la página web. Los motores de búsqueda valoran aquello que afecta a los tiempos de descarga y pueda retrasar la visualización de la información.

Cabe destacar las fuertes dependencias existentes entre determinados factores de posicionamiento. Por ejemplo, la credibilidad de un recurso web es un factor indirecto, que directamente repercute en la popularidad. Asimismo, el número de visitas a un recurso web indica su popularidad, pero también aporta información sobre las preferencias de los usuarios.

2.2 Optimización Web

2.2.1 Introducción

La optimización de páginas para motores de búsqueda, también conocido como *Search Engine Optimization* o SEO, es un conjunto de técnicas aplicadas en el desarrollo web con la finalidad de mejorar la visibilidad de un recurso web (Diligenti et al**.,** 2004), mejorando su posición en el ranking de resultados.

La optimización del posicionamiento web se realiza mediante métodos y técnicas que buscan potenciar, en la fase de diseño, los factores o variables que consideran más importantes los motores de búsqueda.

2.2.2 Optimización en el diseño web

Como consecuencia de las diferencias y modificaciones en los algoritmos de posicionamiento de los motores de búsqueda, los procesos de optimización de un recurso web deben adaptarse a: (1) el motor de búsqueda en el que se pretende ser más visible; (2)

y a la temática y consultas para las que se desea competir por un puesto alto en el ranking de posicionamiento.

2.2.3 Herramientas SEO

Las herramientas SEO son aplicaciones que proporcionan ayuda en la optimización del diseño de páginas web para mejorar su posicionamiento. El éxito de estas herramientas depende de su utilidad para conseguir elevar la posición de una web en los rankings de los buscadores sobre una temática concreta. Estas aplicaciones se basan en el conocimiento que poseen sobre los algoritmos y factores de posicionamiento de los motores de búsqueda para determinar la relevancia web.

Existen múltiples aplicaciones software para facilitar la optimización web. Muchas de estas herramientas tienen carácter propietario, pero también existen una amplia variedad de herramientas disponibles de forma gratuita. A continuación se presentan algunas de las principales herramientas SEO.

2.2.3.1 SEO AddWeb™ Website Promoter 8

La herramienta SEO AddWeb™ Website Promoter 8 pertenece a Cyberspace Headquarters Llc. y aprovecha recursos libres de internet como buscadores web, sitios clasificados, directorios de enlaces, etc. para promocionar los sitios web de sus clientes.

Entre otras funcionalidades proporciona información del tráfico de la Web y de los rankings de los buscadores además de promover el comercio de enlaces entre distintos sitios web.

Al margen de las funcionalidades propias dirigidas a la optimización, dispone de material divulgativo y tutoriales que favorecen el entendimiento de las técnicas empleadas en el diseño web y guían a los usuarios en la aplicación. También ofrece un calculador del retorno de la inversión (ROI). La página web de AddWeb™ Website Promoter 8 es http://www.cyberspacehq.com.

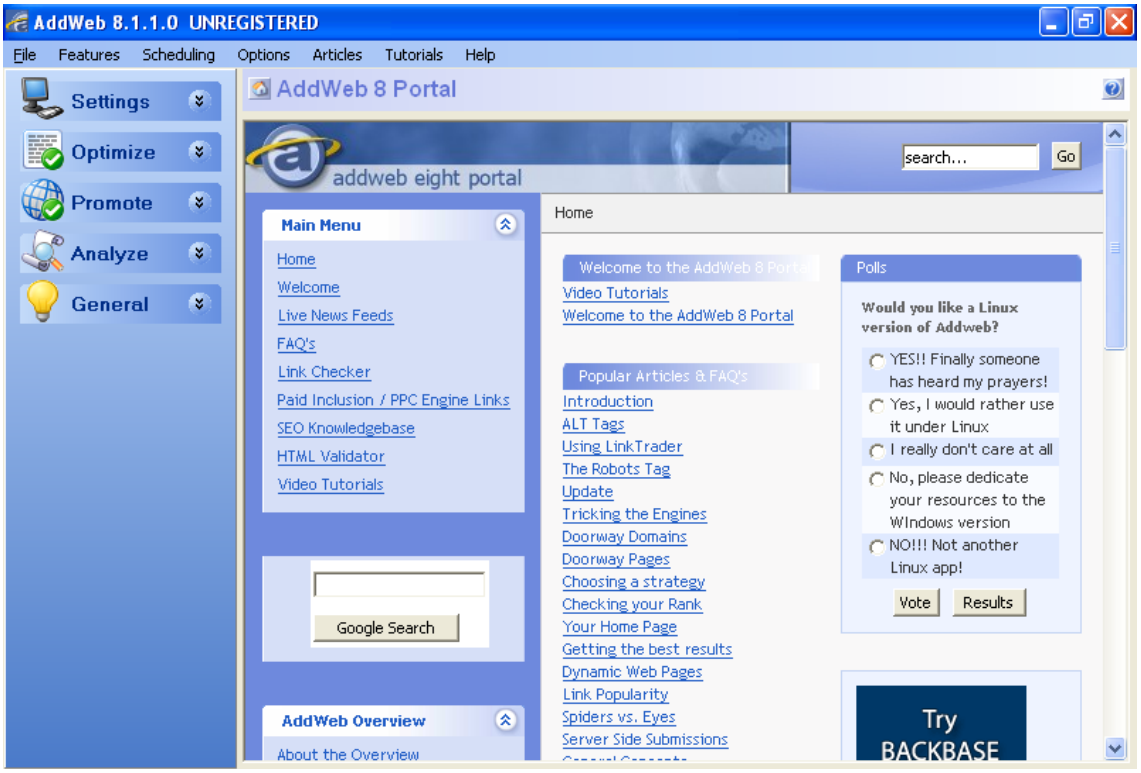


Figura II-4: Interfaz Add Website Promoter

Entre las funcionalidades más interesantes que presenta, destacan algunas como: funcionalidades preventivas de problemas de indexación de las web; funcionalidades específicas de contenido de páginas; funcionalidades de visibilidad promocionada, funcionalidades orientadas al ranking, etc.

En primer lugar, las funcionalidades preventivas de problemas de indexación de las web chequean los enlaces y se comprueba la existencia de posibles enlaces rotos que impidan a los *robots* acceder a las páginas. En este sentido, la aplicación dispone de un validador de código HTML que verifica la correcta construcción de las páginas web para asegurar la catalogación e indexación por parte de los *spiders*.

En segundo lugar existen funcionalidades específicas que revisan el contenido de la página y aconsejan sobre su categorización. De este modo se facilita que las páginas aparezcan en directorios como DMOZ Open Directory con la temática asociada a la categoría.

La aplicación ofrece la posibilidad de intermediar con los motores de búsqueda para una mejora de la visibilidad de los sitios web mediante abono de cantidades (*Paid Inclusion* o *Pay Per Click*). Sin embargo, estas alternativas concretas de mejora del posicionamiento no son consideradas en este trabajo.

El alta o la presentación de un sitio web en un buscador es una funcionalidad que presenta esta herramienta. Contempla la totalidad de los motores de búsqueda del mundo agrupados por países. La aplicación da de alta o presenta las páginas a los buscadores web de forma manual o automática según el caso. Esta última es útil cuando se presenta una web a un gran número de motores de búsqueda por el ahorro de tiempo de conlleva.

Otras funcionalidades están orientadas al análisis del ranking como la posición en el ranking (*ranking position*), el número de páginas del ranking (*ranking page number*), el contador de enlaces de la página (*link count*), o el contador de direcciones URL de la página (*url count*).

La herramienta posee también funcionalidades para determinar las palabras clave más convenientes para el posicionamiento de una web determinada. Para ello, una vez seleccionada una palabra representativa de la página web objeto de optimización, informa del número de búsquedas que se han efectuado en el último mes con dicha palabra. Las palabras que tengan asociados un mayor número de búsquedas serán las más idóneas para presentarlas como palabras clave. Este proceso se puede restringir mediante localización de un país. Se pueden introducir también las páginas web de la competencia para averiguar sus palabras clave. Cuántas menos de estas web compartan las palabras clave del sitio objeto de optimización, más fácil será posicionarlo en puestos preferentes.

La herramienta dispone de un editor de páginas web HTML con versiones textual y gráfica. Además de un constructor de páginas que dirige el diseño con el fin de garantizar la indexación por parte de los spiders.

A continuación se resumen las secciones de aplicación del constructor de páginas de la herramienta AddWeb™ y Website Promoter 8.

SECCIONES DE APLICACIÓN
URL de la página
Enlace de página de inicio, sólo para doorways pages
Page title (*title Tag*): etiqueta de título
Header (*H1 Tag*): etiqueta de cabecera
Palabra clave
Meta description: breve descripción
Body Tag: cuerpo del texto
Comentarios

Tabla II-1: Secciones de aplicación del constructor de páginas de la herramienta AddWebTM y Website Promoter 8

El editor también permite incorporar imágenes con sus correspondientes descripciones (etiquetas ALT) y crear páginas *doorways* de optimización configurable que enlacen a la página principal. La idea de las *doorways* consiste en mostrar un producto óptimo a los robots de indexación y otro a los usuarios que son redirigidos a la verdadera página por un enlace. Resaltar que la utilización de *doorways* puede ser susceptible de penalización si es detectada por los motores de búsqueda.

Otra funcionalidad útil es la comparación de las páginas que se encuentran en las primeras posiciones. Los patrones comunes presentes en estas páginas constituyen pistas sobre las que orientar la optimización. Se puede restringir esta funcionalidad a las páginas que constituyan la competencia directa de la web a optimizar.

La herramienta permite el acceso a un mercado de enlaces con el objetivo de elevar los niveles de popularidad de recursos web. Las páginas orígenes de estos enlaces deben tener una afinidad temática con la web que se desea optimizar para que su influencia sea mayor en el posicionamiento, también se pueden obtener vínculos desde otras páginas mediante *Pay Per Clic* o pago por clic. No obstante, la proliferación excesiva de enlaces puede tener efectos adversos sobre una web si los buscadores la confunden con una granja de enlaces.

Otra funcionalidad relacionada con la popularidad consiste en rastrear el tráfico de una web. Es preciso que previamente se haya incluido en la página código HTML específico. La información obtenida es crucial para encaminar futuras modificaciones en el diseño

web. Por último, la herramienta realiza informes periódicos de los progresos de la optimización del sitio y su estado.

2.2.3.2 Internet Business Promoter 3.0.3

Internet Business Promoter es una herramienta profesional (http://www.Axandra.com). Esta herramienta tiene como objetivo la inclusión en el Top 10 (diez primeras posiciones de los resultados que proporcionan los buscadores) de las páginas que optimiza. Las funcionalidades que ofrece a sus usuarios se centran en tres aspectos principales (Figura II-5):

1. Optimización web dirigida al incremento del número de las visitas de usuarios y por tanto al aumento del volumen de negocio que se publicita, o gestiona con esa web.
2. Presentación de los sitios web a posicionar a los buscadores y directorios en los que desea ser incluidos con el propósito de aparecer en las listas de resultados de los motores de búsqueda.
3. Generación de informes de los resultados obtenidos.

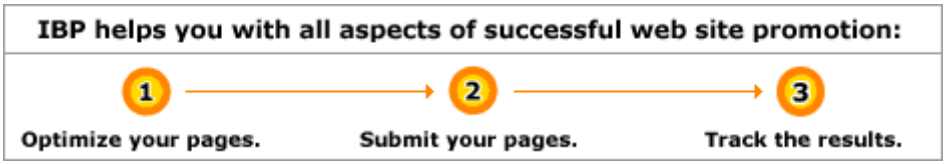


Figura II-5: Aspectos relevantes en la promoción de sitios web

Como se muestra en la Figura II-6, la interfaz se organiza según las funcionalidades de la aplicación.

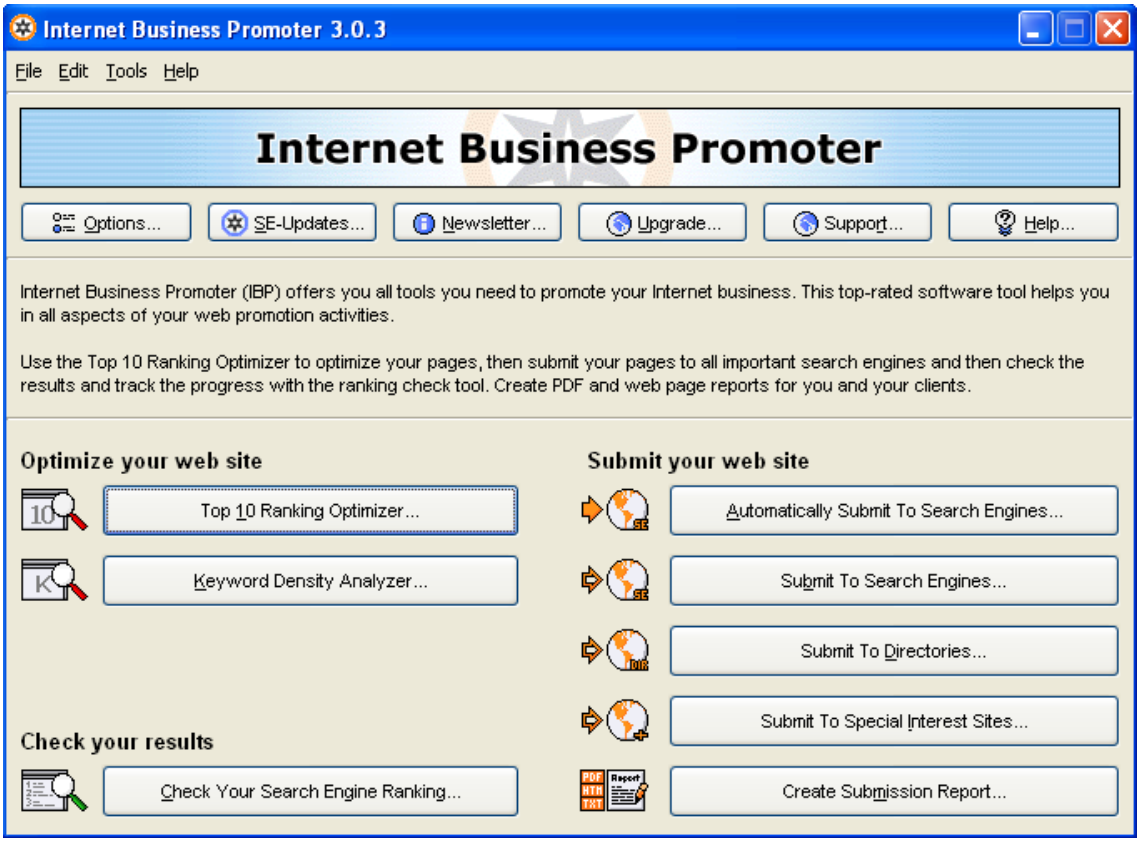


Figura II-6: Interfaz Internet Business Promoter

Se revisan ahora las funcionalidades de la herramienta en el orden recomendado por los autores para maximizar el éxito del proceso de optimización.

La primera recomendación consiste en optimizar las páginas para las palabras clave y motores de búsqueda seleccionados, antes de publicarlas. La estrategia radica en comparar la página a posicionar con los diez primeros resultados de las búsquedas correspondientes a esas selecciones.

Las diferencias existentes entre el código HTML de la página en proceso de optimización y las generalidades encontradas en los códigos de las otras diez indican las posibles deficiencias que hay que solventar.

Los campos analizados por comparación en la búsqueda de patrones para internet de la herramienta Internet Business Promoter se presentan en la siguiente Tabla II-2, dónde se indica que la selección de las palabras clave debe ser evaluada según las siguientes consideraciones:

1. Las palabras clave deben ser representativas de la temática de la página web.
2. La frecuencia de uso por parte de los potenciales usuarios debe de ser elevada.

3. No es conveniente que generen demasiados resultados en las búsquedas.
4. Los autores aconsejan que el número de palabras clave este entre 2 y 4.

CAMPOS DE COMPARACIÓN	
Document title	Título del documento
Meta keywords	Etiqueta de palabras clave.
Meta description	Etiqueta de la descripción
1st Phrase in body	Primera frase del texto del cuerpo del documento
All links text	Texto de todos los enlaces
Link popularity	Popularidad o importancia de los enlaces.
All links URL	Todas las URL de los enlaces
Same site link URL	URL de los enlaces de página similares.
Outbounds links URL	URL de los enlaces externos.
<H1> y <H2> headline texts	Texto de las cabeceras H1 y H2
Same site link texts	Textos de los enlaces de sitios similares
Outbound link texts	Textos de los enlaces externos.
HTML comments	Comentarios de HTML
Image Alt attributes	Características de las etiquetas ALT de las imágenes

Tabla II-2: Campos de comparación en la búsqueda de patrones para Internet Business Promoter

Para búsqueda de palabras claves apropiadas la aplicación deriva a la herramienta SEO Word Tracker que se analizará más adelante. Esta herramienta sí que estudia directamente la densidad de las palabras clave de las páginas web, incluso de las que aún no se han publicado. Los campos de análisis son los siguientes: titulo del documento, texto del cuerpo del documento, etiqueta de palabras clave, todas las URL de los enlaces, URL de los enlaces de página similares, URL de los enlaces externos, características de las etiquetas ALT (imágenes), combinación de las partes de la página web, texto en negrita en el cuerpo del documento, etiqueta de la descripción, texto de todos los enlaces, textos de los enlaces de sitios similares, etc.

El listado completo se encuentra en la siguiente Tabla II-3.

SECCIONES DENSIDAD DE PALABRAS CLAVE	
Document title	Titulo del documento
Body text	Texto del cuerpo del documento
Meta keywords	Etiqueta de palabras clave
All links URL	Todas las URL de los enlaces
Same site link URL	URL de los enlaces de página similares
Outbounds links URL	URL de los enlaces externos
Image Alt attributes	Características de las etiquetas ALT (imágenes)
Web Page Parts Combined	Combinación de las partes de la página web
Bold body text	Texto en negrita en el cuerpo del documento.
Meta descripción	Etiqueta de la descripción
All links text	Texto de todos los enlaces
Same site link texts	Textos de los enlaces de sitios similares
Outbound link texts	Textos de los enlaces externos.
HTML comments	Comentarios de HTML

Tabla II-3: Secciones de densidad de palabras clave en Internet Business Promoter

Las funcionalidades de la herramienta asociadas a la presentación de páginas web a los buscadores y directorios son automáticas o semiautomáticas, se pueden simultanear y requieren la información de la categoría temática.

También dispone la herramienta de funcionalidades relativas a las posiciones que ocupan en los rankings las páginas web siempre que se seleccionen palabras clave y buscadores.

Por último, la aplicación emite informes para todas sus funcionalidades proporcionando resultados e indicando el estado en que se encuentran los procesos. Mediante suscripción se puede obtener información novedosa sobre posicionamiento web y métodos de optimización. De esta forma, los propietarios de la web pueden saber de primera mano cómo afectarán las modificaciones que surjan en el posicionamiento de sus páginas.

2.2.3.3 FlashMarketing's Spider 1.86

La herramienta SEO FlashMarketing's Spider 1.86[8] (www.flashmarketing.com/spider) está dirigida a usuarios inexpertos en el mundo web. Su cometido es la presentación automática de sitios web a buscadores, directorios, páginas de enlaces, clasificados y otros recursos en línea que favorezcan su visibilidad.

La sencillez de su interfaz permite introducir de forma intuitiva los datos del sitio web que se desea presentar. La inclusión de estos datos es estándar y por tanto válida para todas las presentaciones.

Figura II-7: Interfaz Flash Marketing's Spider

El programa, según los autores, debe ejecutarse en el orden en que se muestran las pestañas de la parte superior de su interfaz. Después de introducir los datos identificativos relativos a la web a presentar se solicita, entre otra información, las palabras clave (hasta dieciséis) y la categoría temática. Si existen dudas sobre la categoría la aplicación es capaz de asignarle una automáticamente. Los procesos de presentación se siguen mediante mensajes de confirmación e históricos de ejecuciones anteriores.

2.2.3.4 Web position platinum 3.5

La herramienta *Web Position Platinum 3.5*[9] (http://www.webposition.com/) ofrece, en el orden óptimo de aplicación, un conjunto de funcionalidades para mejorar el posicionamiento de páginas web. Las funcionalidades y el orden de aplicación aconsejado se muestran en la siguiente Figura II-8:

[8] Copyright © 1997-1999 PBD Technologies Traffic Builder
[9] Copyright © 1997-2006 Web Trenes Inc.

Figura II-8: Esquema Web Position

La interfaz de inicio de esta herramienta SEO se muestra en la siguiente Figura II-9 y presenta cinco funcionalidades: rankings, palabras clave, construcción y optimización de páginas web, presentación de páginas y rastreo de resultados.

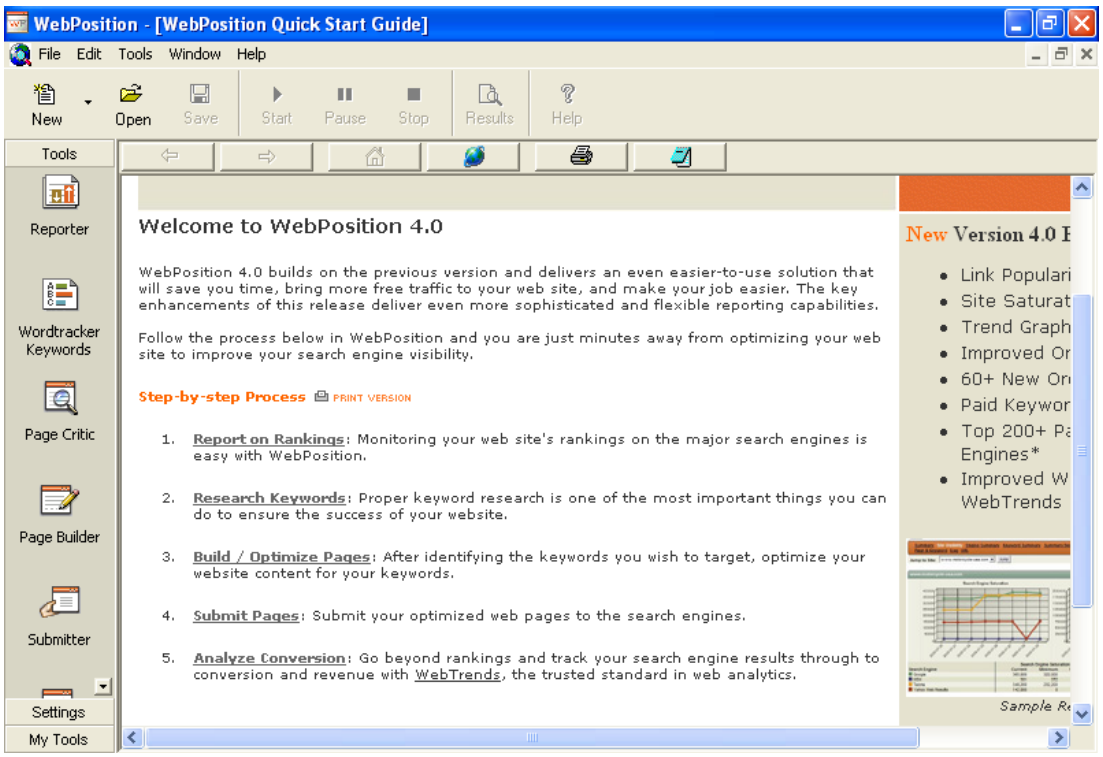


Figura II-9: Interfaz herramienta SEO Web Position

1. **Informe de Rankings**: En el estudio de rankings se puede incluir la página propia o las de la competencia. Las palabras clave pueden ser importadas de la propia web o seleccionadas de forma independiente para poder jugar con otras

posibilidades. Si alguna página no aparece en los rankings la herramienta puede comprobar si se trata de un problema de indexación, incluso en el caso de que estas páginas se encuentren en niveles muy inferiores del dominio. Los resultados se muestran en gráficas e informes detallados e intuitivos. Se ofrece un resumen principal y una serie de aspectos presentes en la siguiente Tabla II-4:

INFORMACIÓN OBTENIDA DEL ANÁLISIS DE RANKING	
Summary	Resumen de las estadísticas de los rankings
Visibility	Visibilidad de la página web, incluye graficas, número de páginas indexadas del dominio por buscador y popularidad del sitio por sus enlaces
Engine	Posición de la página para cada palabra clave agrupando los resultados por buscador
Keyword	Posición de la página en cada motor de búsqueda agrupando los resultados por palabras clave
Listings	Ofrece los resultados de las palabras clave en los motores de búsqueda
Detail	Ofrece los resultados de las palabras clave de cada motor de búsqueda por separado
Alert	Alerta de errores producidos en la petición de datos de los buscadores
Trend	Muestra las graficas de los cambios de posición de la página con cada palabra clave en cada buscador
Competitive	Muestra las posiciones respecto a la competencia con las palabras clave seleccionadas
URL/Keyword	Muestra las posiciones para cada palabra clave junto a las URL del dominio que las alcanzan
Log	Registro de los detalles de la ejecución, búsquedas completadas y fallidas
URL	Muestra el estado de la URL tras su verificación

Tabla II-4: Resultados obtenidos del estudio de los rankings

2. **Búsqueda de palabras clave.** Se proponen 5 puntos teóricos en los que fundamentar la elección de las palabras clave:

    i. El selector de palabras clave debe pensar como un usuario que demanda información relacionada con la temática de la página en diseño.

ii. Se deben orientar las palabras clave a aquellas que son propias del nicho de mercado al que se dirigen los productos de la web en construcción.

iii. Generar múltiples opciones sin reflexionar demasiado (tormenta de ideas) que luego serán filtradas por su idoneidad. Las palabras clave de la competencia pueden servir como embrión del proceso.

iv. Seleccionar únicamente las palabras clave relevantes y significativas de la página web a optimizar.

v. Evitar las palabras clave polisémicas.

Con el fin de analizar lo óptimas que serán las palabras claves seleccionadas con estas estrategias se recomienda el uso el uso de la herramienta WORDTRACKER que está embebida en la aplicación.

3. **Construye y optimiza las páginas web**. La optimización se realiza en base a características comunes encontradas entre las web mejor posicionadas en los rankings en los que se desea una mejor visibilidad.

CAMPOS DE COMPARACIÓN	
Title	Titulo
Meta Keyword	Etiqueta de palabras clave
Meta descripción	Etiqueta de descripción de la página.
Heading	Cabecera
Link text	Texto del enlace
Hiperlink text	Texto del hipervínculo
Alt	Etiquetas de las imágenes
Commente	Comentarios
Body text	Texto del cuerpo del documento

Tabla II-5: Campos de comparación estudiados en los primeros resultados de las consultas

Se pueden seleccionar los campos a comparar entre los que se presentan en la Tabla II-5.

La herramienta dispone de un editor de código HTML para aplicar directamente las modificaciones aconsejadas que otorgarían un mejor posicionamiento.

Otros aspectos interesantes de la información que proporciona la herramienta están relacionados con ciertas propiedades de las páginas web que son valoradas o penalizadas por los buscadores. El editor comprueba estas propiedades que se listan en la Tabla II-6:

CAMPOS DE COMPROBACIÓN DEL EDITOR
Número máximo de repeticiones de la palabra clave en una fila
¿Ha usado el mismo color para el texto y el fondo?
¿Tiene áreas de entrada ocultas?
¿Usa la etiqueta Meta refresh?
¿Usa Frames?
¿Usa controles?
¿Usa Javascript?
¿Usa VBScript?

Tabla II-6: Campos de comprobación de Web Position

4. **Presentación de páginas**: Se ofrece la posibilidad de presentar la página a los índices de los directorios y buscadores del mercado para que incluyan la página y sea visible en los resultados. La herramienta no da importancia a que el contenido no sea definitivo pudiéndose dejar parte del trabajo para futuras actualizaciones. Esta funcionalidad de presentación de páginas web no parece muy cuidada ya que los buscadores web suelen solicitar más información que la que pide la herramienta.

5. **Rastreo de resultados:** Para finalizar, se verifica que se ha llevado a cabo la optimización de una manera eficiente. Se aconseja el uso de la herramienta de análisis "WEB TRENDS", que se estudiará posteriormente, para descubrir tendencias y novedades a tener en cuenta en el mantenimiento de las web.

2.2.3.5 Web CEO Version 6.0

La herramienta SEO Web CEO Version 6.0[10] (http://www.webceo.com/) se estructura en cuatro módulos y 12 funcionalidades relacionadas con el análisis del tráfico web, y la promoción y mantenimiento de páginas web.

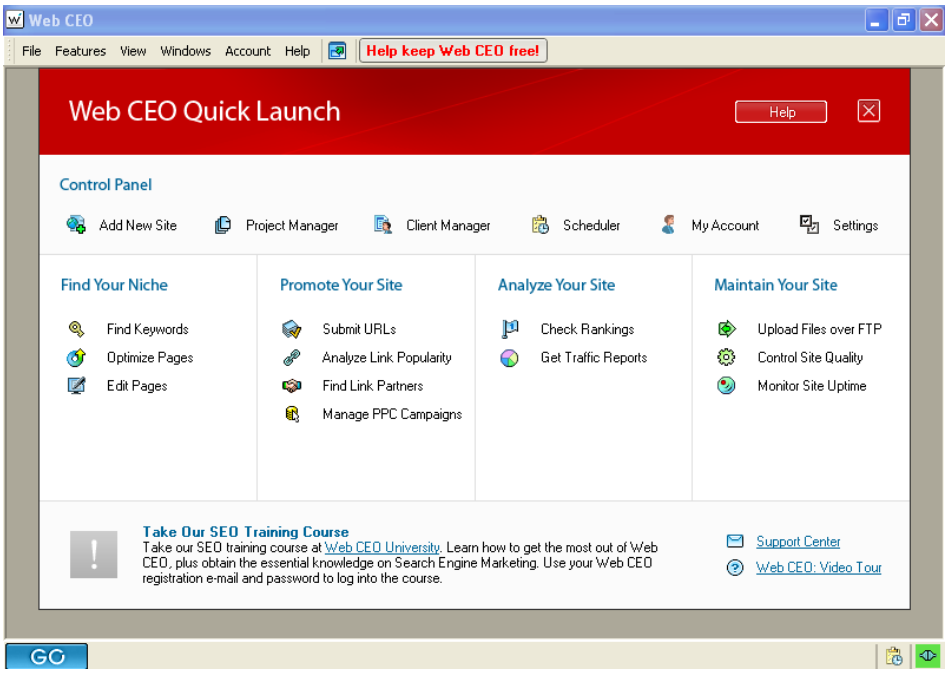


Figura II-10: Interfaz herramienta Web CEO 6.0

Esta aplicación efectúa análisis de palabras clave, sobre los buscadores Google, Yahoo Search y MSN. Selecciona aquellas palabras clave que describen el negocio asociado a la web a optimizar. Este análisis se basa en comparaciones de las páginas que se encuentran en las primeras posiciones de los rankings, relativos a las palabras en estudio en los citados buscadores.

Las comparaciones determinan las palabras clave más populares a partir de estudios de frecuencias. El grado de competencia se calcula por el número de páginas que contienen estas palabras en sus METAKEYWORD.

El conjunto de características sobre el que se obtiene información tras el análisis de palabras clave en las páginas de la competencia se muestra en la siguiente Tabla II-7.

[10] Web CEO Ltd.

CARACTERÍSTICAS DEL ANÁLISIS POR PALABRA CLAVE EN LA COMPETENCIA	
Daily World Searches	Búsquedas diarias en el mundo con la palabra clave seleccionada. A mayor nº más idoneidad de las palabras
Pages with keywords	Páginas rankeadas por Google que contienen la palabra clave. A menor nº menos competencia
Titles with keywords	Resultados de Google con la palabra clave en su titulo. Es un indicador de las páginas que han optimizado para el término en cuestión, cuantas menos mejor
Links to #1 (2,…,10)	Popularidad de los enlaces. Si las 10 primeras web tienen un gran nº de habrá que incrementar este factor de popularidad en la página a optimizar
Traffic Rank for #1 (2,..,10)	El ranking del tráfico (basado en Alexa) A menor ranking mayor popularidad
Bid #1 (2,…,10)	Muestra los enlaces patrocinados en Yahoo Search. Si los enlaces de la competencia son de pago los usuarios intuyen que sus posiciones privilegiadas no se deben a su calidad
PR #1 (2,…,10)	PageRank de Google. Si el PageRank de los primeros puestos es 5 o mayor puede ser difícil posicionarse entre ellos

Tabla II-7: Características del análisis por palabra clave en la competencia

Después de seleccionar las palabras clave se procede a la optimización mediante diseño web. En este sentido el análisis que realiza esta herramienta SEO es de los más completos que se han estudiado en esta investigación. Las características que analiza se estructuran en la siguiente Tabla II-8:

CARACTERÍSTICAS ANALIZADAS		
	General page properties:	**Propiedades generales de la página**
Page	*HTML size (Kb)*	Tamaño del código HTML
	Last modified	Ultima modificación
	Has same color text and background	Igual color de texto que el fondo
	Has tiny text	Texto muy pequeño
	Has immediate keyword repeats	Palabras clave consecutivas repetidas
	Uses controls	Uso de controles
	Uses frames	Uso de Frames

	Uses external JavaScript	Uso de JavaScript externo
	Uses internal JavaScript	Uso de JavaScript interno
	Uses external VBScript	Uso de VBScript externo
	Uses internal VBScript	Uso de VBScript externo
	File robots.txt disallows spidering	Uso de fichero que no permite ser Indexado
Page URL	*Keyword as a part of URL (domain, folder and page name)*	Palabra clave parte de URL
	Keyword as a separate part of URL (domain, folder and page name)	Palabra clave como parte separada de la URL
	Main on-the-page factors influencing your ranking	**Principales factores que influyen en el ranking**
<HEAD>	**<TITLE>**	
	Number of Titles	Nº de títulos
	First tag in the <HEAD> tag	Primera etiqueta en las cabeceras
	Characters in Title	Nº Caracteres del título
	Words in Title	Nº Palabras del título
	Stop words in Title	Palabras en titulo
	Keyword frequency in Title	Frecuencia de palabras clave en titulo
	Keyword prominence in Title	Importancia de la palabra en titulo
	Number of Titles	Nº de títulos
	META Description	
	Number of META Description tags	Nº de etiquetas de descripción
	Characters in META Description	Nº Caracteres en la descripción
	Words in META Description	Nº de palabras en la descripción
	Stop words in META Description	Palabras en la descripción
	Keyword frequency in META Description	Frecuencia palabras clave en el título
	Keyword prominence in META Description	Importancia de la palabra en la descripción
	Keyword weight in META Description	Peso de la palabra clave en la descripción

	META Keywords	
	Characters in META Keywords tag	Nº caracteres en etiqueta de palabras clave
	Number of META Keywords tags	Nº de etiquetas de palabras clave
	Words in META Keywords tag	Nº palabras en palabras clave
	Keyword frequency in META Keywords	Frecuencia de palabras clave en la etiqueta.
	Keyword prominence in META Keywords	Importancia de la palabra clave
	Keyword weight in META Keywords	Peso de la palabra clave
	META Refresh	
	Refresh	Actualizar
	Refresh time	Tiempo de Refresh
	Redirect	Redirección
	META Robots	
	None	Nulo
	No index	No indización
	No follow	No seguir el enlace
	No archive	No archivar
<BODY>	*Visible text*	Visibilidad del texto
	Words in Body	Palabras en el cuerpo de la página
	Bold keywords in Body	Palabras clave en negrita en el cuerpo
	Underlined keywords in Body	Palabras clave subrayadas en el cuerpo
	Keyword frequency in Body	Frecuencia de las palabras clave en texto
	Keyword prominence in Body	Importancia de las palabras clave
	Keyword weight in Body	Peso de las palabras clave
	Keyword at the beginning of Body	Palabra clave en el inicio del cuerpo
	Keyword at the end of Body	Palabra clave en el final del cuerpo
	First heading on the page (H1-H6)	Primeras Cabeceras
	Keyword frequency	Frecuencia de la palabra clave

	Keyword prominence	Importancia de la palabra clave
	Keyword weight	Peso de la palabra clave
	All headings	Todas las cabeceras
	Headings on the page	Nº Cabeceras en la página
	Headings with the keyword	Porcentaje cabeceras con palabra clave
	Keyword frequency in all headings	Frecuencia palabra clave en todas las cabeceras
	Keyword weight in all headings	Peso de las palabras clave en cabecera
	Links	Enlaces
	Total links on the page	Nº total de enlaces
	Links to the external pages	Enlaces a páginas externas
	Text in links including ALTs	Texto de los enlaces incluido en las etiquetas ALT
	Links with keyword in the text and ALT	Enlaces con la palabra clave en el texto
	Keyword frequency in links (text and ALT)	Frecuencia de aparición de la palabra clave en el texto
	Keyword weight in links (text and ALT)	Peso de la palabra clave en el texto del enlace
	ALT image attributes	Atributos de las imágenes ALT
	ALT attributes on the page	Nº atributos en la página.
	ALT attributes with the keyword	Nº atributos con palabras clave
	Keyword matches in the first 3 ALT attributes	Palabra clave en los 3 primeros atributos ALT
	Keyword frequency in ALT attributes	Frecuencia de la palabra clave en los atributos
	Keyword weight in ALT attributes	Peso palabra clave en los atributos ALT
	Comments	Comentarios
	Words in comments	Nº Palabras en los comentarios
	Keyword frequency in comments	Frecuencia palabra clave en comentarios
	Keyword weight in comment	Peso palabra clave en los comentarios

Tabla II-8: Características analizadas Web CEO 6.0

La aplicación también dispone de un editor de páginas web para facilitar las tareas de optimización que implican un rediseño de las páginas. Se muestran en la siguiente Tabla II-9, las opciones de optimización del editor:

CAMPOS DE OPTIMIZACIÓN DEL EDITOR	
Setup	Configuración
Meta Tags	Etiquetas meta
Headings	Cabeceras
Alt	Imágenes
Links	Enlaces
Content	Contenido
Opcional Meta Tags	Etiquetas META opcionales
Find & replace	Encuentra y reemplaza

Tabla II-9: Campos de optimización Web CEO 6.0

Al igual que en otras herramientas, las tareas de indexación se realizan mediante protocolos de presentación automáticos o manuales. En estas presentaciones se seleccionan las páginas del dominio que se desean indexar y los buscadores web y directorios, generándose luego un informe con el estado del proceso.

Por otra parte, esta herramienta tiene una alta capacidad para analizar y gestionar enlaces siendo por tanto de gran utilidad para el incremento de popularidad de las web. La aplicación informa mediante gráficas e históricos del número de páginas que tienen enlaces hacia la página objeto de promoción, se pueden incluir también las páginas de la competencia con el fin de realizar análisis comparativos. Además, se pueden explorar las páginas de las que parten los vínculos con el propósito de averiguar si son de una temática afín.

En cuanto a la gestión de enlaces, la filosofía de la herramienta es obtener enlaces por medio de acuerdos entre páginas. Para ello busca potenciales "socios" de enlaces en la web, preferentemente páginas que compartan una misma categoría temática. Las comunicaciones para la negociación y establecimiento de acuerdos se mantienen por un gestor vía mail. También contempla esta herramienta la posibilidad de obtener enlaces por acuerdos de *Pay Per Clic* (PPC).

Otra medida de la popularidad proviene de las visitas de los usuarios a las páginas web, en este sentido esta herramienta lleva a cabo un registro del tráfico recibido en un sitio web concreto.

Como se muestra en la siguiente Figura II-11 es necesaria la inclusión de código HTML en la página web para registrar el acceso de los usuarios. Con esta información se descubren las preferencias de los usuarios y el interés que tienen en el contenido de las web.

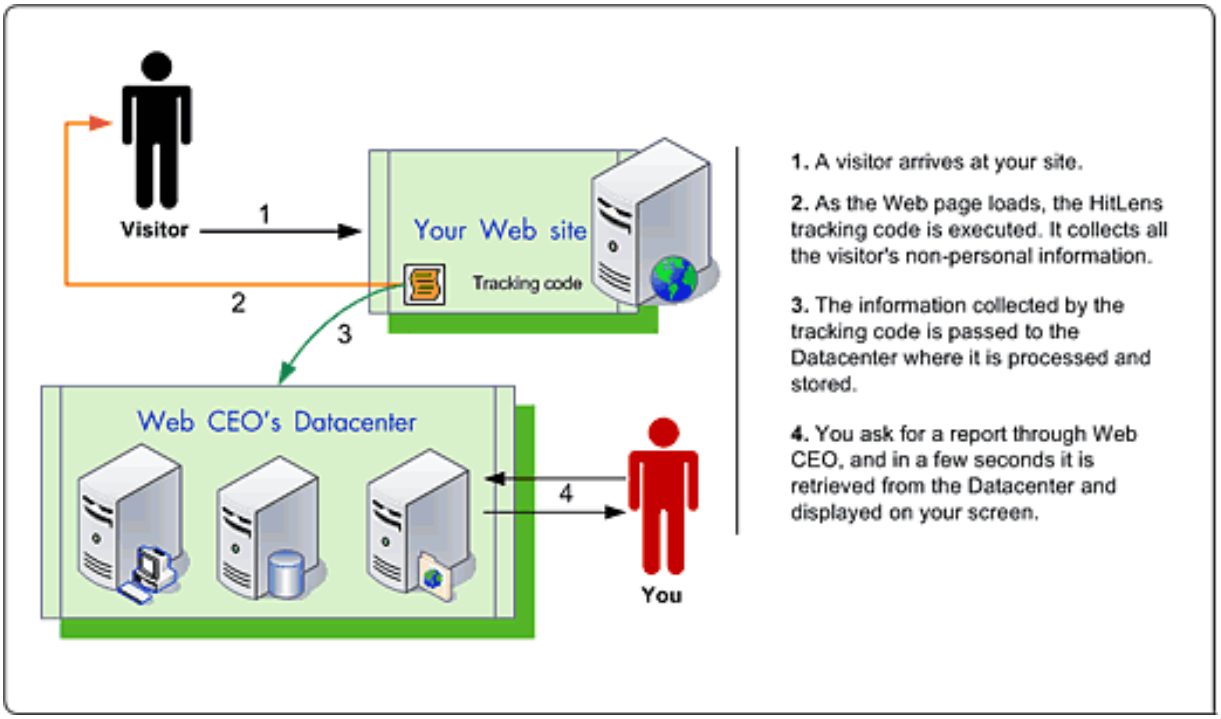


Figura II-11: Método de rastreo en Web CEO 6.0

Las tareas de mantenimiento se realizan mediante protocolos de actualización de páginas web previo chequeo de posibles errores. Las posibles opciones al habilitar el chequeo de errores se muestran a continuación (Tabla II-10). Como se muestra en la tabla, destacan los enlaces rotos, la falta de imágenes, las páginas sin título, anclajes rotos, carencia de etiquetas meta, contenidos antiguos y páginas lentas.

La herramienta ofrece una funcionalidad interesante relacionada con la accesibilidad web. Realiza un test de velocidad de descarga de las páginas web y mide la velocidad de descarga del sitio desde Estados Unidos de América, el Reino Unido o desde el ordenador de la monitorización. La unidad de medida utilizada son bytes por segundo.

CHEQUEO DE ERRORES	
Broken links	Enlaces rotos
Missing images	Imágenes que faltan
Untitled pages	Páginas sin titulo
Slow pages	Páginas lentas
Broken anchors	Anclajes rotos
Missing Meta tags	Faltan etiquetas Meta
Stale content	Contenido antiguos

Tabla II-10: Chequeo de errores en Web CEO 6.0

Finalmente, se presenta un informe con todas las páginas incluidas en el sitio web mostrando los problemas de cada página. Si los tiempos no son adecuados conviene rediseñar las páginas, de lo contrario es posible que los problemas de accesibilidad desmotiven al usuario y no se efectúe la visita.

Figura II-12: Test de velocidad en Web CEO 6.0

2.2.3.6 SEO Tools ™

Las herramientas SEO Tools ™[11] (http://www.seochat.com/seo-tools/) son un conjunto de aplicaciones que facilitan el diseño de las páginas web con la intención de generar tráfico y alcanzar puestos preferentes en los rankings de resultados de los buscadores. Se van a describir únicamente las herramientas que son de utilidad en la optimización del diseño de los sitios web.

[11] © 2001-2006. All rights reserved. SEO Chat Cluster 3 hosted by Hostway

Advanced Meta-Tags Generator Tool © SEO Chat™

El propósito de esta herramienta es crear en código HTML las etiquetas META de las páginas web para garantizar una correcta indexación. La herramienta ofrece los campos necesarios para las etiquetas dando una breve descripción de cómo rellenar cada una de ellas. Los campos de la aplicación se muestran a continuación en la Tabla II-11:

CAMPOS A ETIQUETAR	
Title	Título
Author	Autor
Subject	Tema de la página web
Description	Descripción
Classification	Clasificación o categoría de la página web
Keywords	Palabras clave de la página web
Geography	Punto geográfico
Language	Idioma
Expires	Caducidad de la página web
Cache Control	Nivel de control de cache
No Cache	No Cache
Copyright	Propietario de la licencia
Zip Code	Código postal
City	Ciudad
Country	País
Designer	Diseñador
Publisher	Publicista
Revisit-After	Tiempo de actualización de la página web
Distribution	Distribución
Robots, Spiders	Arañas
MS Tags	Etiquetas de Microsoft

Tabla II-11: Campos que etiqueta Advanced Meta-Tags Generator Tool

Una vez incluidos estos datos se genera el código HTML correspondiente.

Alexa Rank Comparison Tool © SEO Chat™

Esta herramienta ofrece un histórico de datos del trafico para cada una de las web basado en las visitas realizadas por los usuarios que tienen instalada la barra de herramientas o *toolbar* de Alexa.

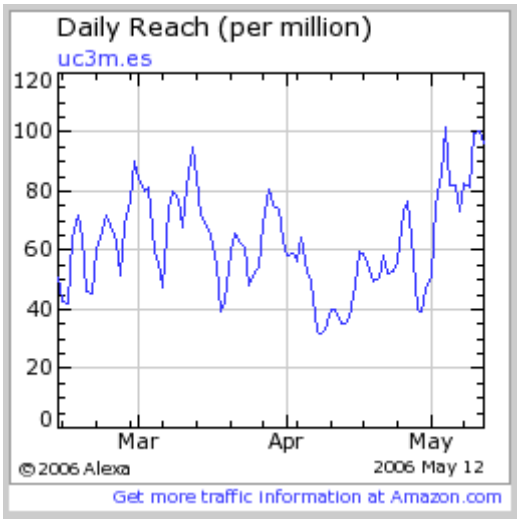


Figura II-13: Gráfico del tráfico en Alexa

SEO Tools – Class C checker

La herramienta *Class C checker* realiza un chequeo de las direcciones de red de distintos dominios devolviendo las direcciones IP y las máscaras que indican las direcciones de red. La finalidad es detectar los enlaces entre direcciones de una misma red ya que pueden ser poco valorados por los buscadores o peor aún, considerar que se ha aplicado una técnica engañosa.

Results for:

Host	IP	Class C
www.ucm.es	147.96.1.15	147.96.1
www.uc3m.es	163.117.136.249	163.117.136

Figura II-14: Resultados de direcciones de red Class C Checker

SEO Tools – Code to Text Ratio

La herramienta *Code to Text Ratio* calcula el porcentaje de texto del documento respecto al total de líneas del código de la página web de HTML. Algunos motores de búsqueda tiene en cuenta este factor porque es un indicador de la cantidad de contenido.

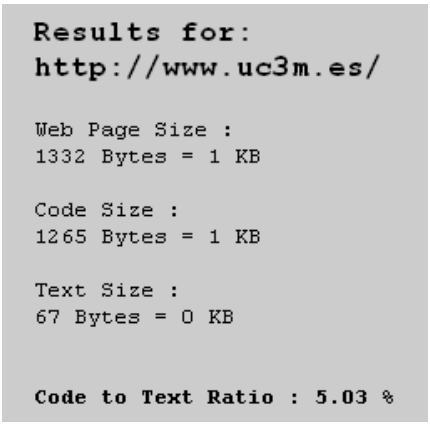


Figura II-15: Resultados porcentaje de texto Code to Text Ratio

No obstante, aunque algunos expertos en herramientas SEO consideran que las páginas web con mucho código y poco texto son poco relevantes para los buscadores, la experiencia indica que el posicionamiento lo realizan únicamente en relación al texto (López, 2009).

SEO Tools – Domain Age

La herramienta *Domain Age* informa sobre la edad o antigüedad de las páginas web. Las web más novedosas o las que demuestran una larga permanencia suelen obtener una mejor valoración.

SEO Tools – Future PageRank

La herramienta *Future PageRank* a partir de los datos de diversos servidores de datos de Google estima el PageRank que obtendrá una web a corto plazo.

SEO Tools – Keyword Suggestions for Google

La aplicación *Keyword Suggestions for Google* sugiere candidatos a palabras claves a partir de términos que se introducen en la herramienta por considerarse descriptivos para determinada página web.

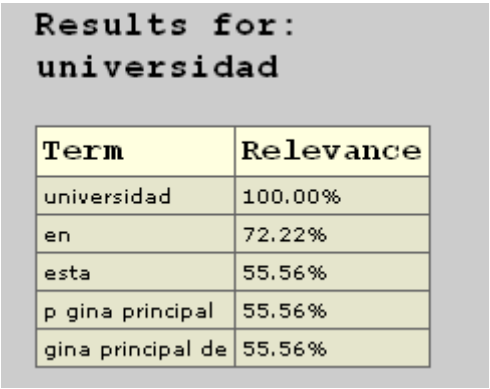
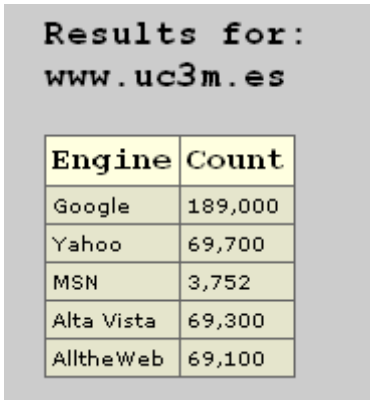


Figura II-16: Resultados palabras clave en Keyword Suggestions for Google

Las sugerencias se basan en un menor número de competidores y se acompañan del porcentaje de relevancia para Google.

SEO Tools – SEO Tools - Indexed Pages

La herramienta *Indexed Pages* obtiene el número de páginas pertenecientes a un sitio Web que han sido indexadas por los motores de búsqueda más conocidos (Google, MSN, Yahoo Search, Alta Vista).



Results for:
www.uc3m.es

Engine	Count
Google	189,000
Yahoo	69,700
MSN	3,752
Alta Vista	69,300
AlltheWeb	69,100

Figura II-17: Resultados indización de páginas en Indexed Pages

Introduciendo la URL del sitio web objeto del estudio se obtiene una tabla similar a la de la Figura II-17.

SEO Tools – Keyword Cloud

La herramienta *Keyword Cloud* efectúa recuentos de palabras clave en las zonas del texto correspondientes a titulares o en las que la caja de letra destaca por su mayor tamaño.

SEO Tools – Keyword Density Tool

El servicio que presta la herramienta *Keyword Density Tool* consiste en analizar la densidad de aparición de las palabras de una página web. Entre las opciones para los análisis se pueden incluir distintas secciones de los documentos como las de las etiquetas META o ALT. En los resultados que se obtiene se presentan las palabras con el número de apariciones y el porcentaje de aparición.

SEO Tools – Keyword Difficult Check

La dificultad para posicionar una página web para determinada palabra clave o frase depende del número de búsquedas que se realizan con esos términos y de la competencia, es decir, del número de páginas que contienen esas mismas palabras clave. Basándose en esos datos, la herramienta *Keyword Difficult Check* estima la dificultad para posicionar una web en la primera posición del ranking de resultados obtenidos al consultar por esos términos. La medida se da en porcentaje tal como ve en la siguiente Figura II-18.

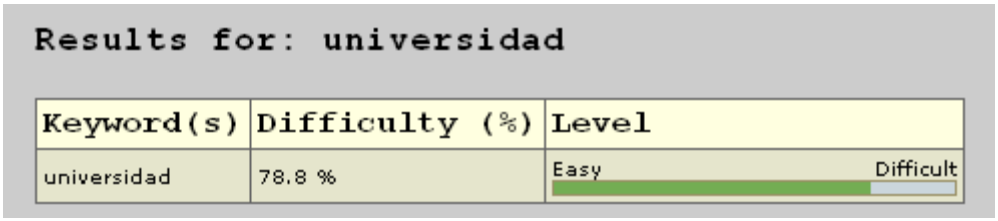


Figura II-18: Estimación para posiciona en Kewword difficult Check

SEO Tools – Multiple Datacenter Keyword Position Check

La herramienta *Multiple Datacenter Keyword Position Check* solicita a los diferentes servidores de datos de Google la posición que ocupa una web en cada uno de ellos para descubrir pautas de optimización. Los datos de entrada son la URL del sitio web y la palabra clave. La salida corresponde a las posiciones para cada dirección de los servidores.

SEO Tools – Keyword Typo Generator

La herramienta *Keyword Typo Generator* aprovecha los errores más comunes que cometen los usuarios al introducir las palabras clave, para sugerir las equivocaciones más frecuentes como alternativa a estas palabras. De esta forma es posible que se elimine parte de la competencia y se obtenga un mejor posicionamiento.

SEO Tools – Link Popularity

La herramienta *Link Popularity* efectúa un recuento de los enlaces dirigidos a una web en los motores de búsqueda Google, Yahoo Search, MSN y Teoma. Con esta información se puede tener cierto control sobre la popularidad de una web.

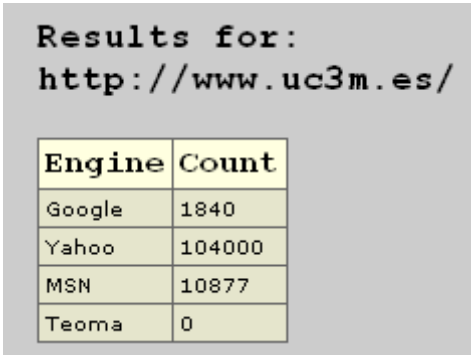


Figura II-19: Resultados popularidad de un enlace en Link Popularity

SEO Tools – Meta Analyzer

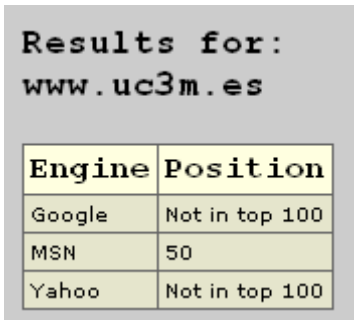
Con el fin de facilitar el estudio en zonas específicas de un documento web, la herramienta *Meta Analyzer* desglosa las etiquetas meta. Las áreas que se independizan para los análisis son Título, Descripción y Palabras Clave.

SEO Tools – PageRank Search.

La herramienta *PageRank Search* actúa de la misma forma que el buscador Google con la única diferencia de que acompaña a cada resultado, fruto de una consulta, con el valor del PageRank correspondiente. Es útil para conocer el PageRank de las páginas con las que se compite en las palabras clave.

SEO Tools – Search Engine Keyword Position

La herramienta *Search Engine Keyword Position* muestra la posición de una web en los rankings obtenidos al consultar con una palabra clave o frase en los buscadores Google, Yahoo Search y MSN. De esta forma se puede conocer lo optimizada que está una página web con respecto a determinadas palabras clave. Tal como se muestra en la Figura II-20 no se especifica la posición exacta de una página web si no está comprendida entre los cien primeros resultados de la consulta realizada.



Results for:
www.uc3m.es

Engine	Position
Google	Not in top 100
MSN	50
Yahoo	Not in top 100

Figura II-20: Resultado Search Engine Keyword Position

SEO Tools – Spider Simulator

La herramienta *Spider Simulator* simula la visión de una araña, de esta forma se pueden tener conocimiento de cómo se indexaran las páginas al ser encontradas por los *robots* que recorren la Web. Así se pueden evitar lecturas inadecuadas que impiden una óptima indexación.

2.2.3.7 Search Engine Commando™ version 3.1.0.440

La herramienta *Search Engine Commando™ version 3.1.0.440*[12] (www.searchengine commando.com/) está concebida para ayudar a los diseñadores web en el incremento del tráfico de sus páginas mediante funciones de promoción.

[12] Search Engine Commando, LLC.© Derechos de autor 2002 Tates Creek Software, LLC

Los autores apuestan sobre los resultados obtenidos afirmando que tras su uso las páginas optimizadas se posicionarán en el Top 10 de los principales motores de búsqueda. La aplicación presenta una interfaz amigable con la que se accede a varios módulos.

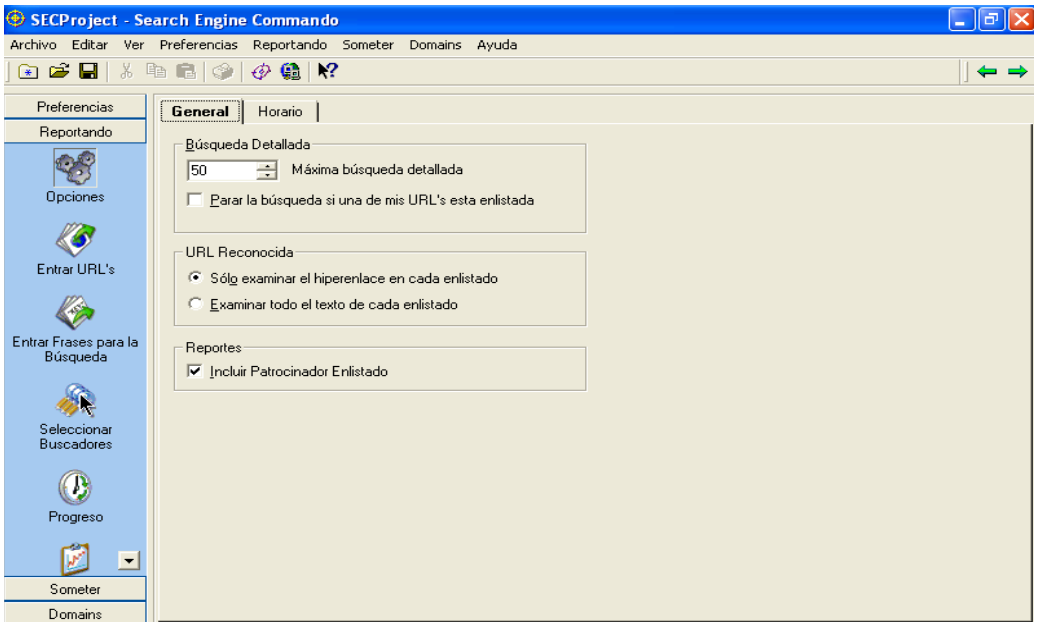


Figura II-21: Interfaz proyecto SEO en Search Engine Commando

La herramienta permite seleccionar el idioma siendo el español una de las opciones. Una vez configurado el idioma se pueden elegir términos de búsqueda u optimización para realizar estudios de ranking en los principales motores de búsqueda del mundo.

Las presentaciones de las páginas a los motores de búsqueda se llevan a cabo tras introducir una serie de datos identificativos y las categorías temáticas de las páginas web. Como resultado se obtiene un listado de incidencias que indica el éxito de la presentación.

La herramienta también permite monitorizar los dominios web con el fin de comprobar su validez. Las comprobaciones se realizan sobre los datos de sus autores, su fecha de creación, su fecha de caducidad, o la existencia del dominio.

2.2.3.8 Agente Web Ranking 2.6.3 Profesional Edition

La herramienta SEO *Agente Web Ranking*[13] (www.agentwebranking.com/) automatiza la verificación de posiciones de las páginas web en los principales motores de búsqueda del mercado.

[13] Copyright 1998-2005 AgentWebRanking

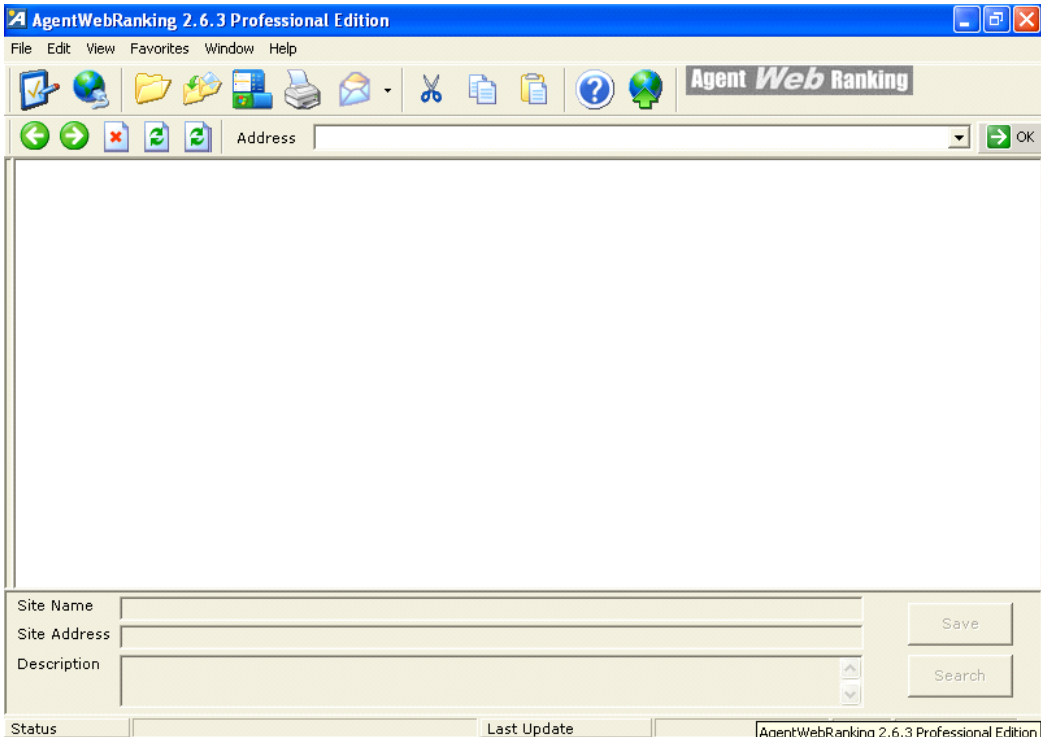


Figura II-22: Interfaz de Agent Web Ranking

Se requiere la elección previa de las palabras clave y del buscador web para los que se desea optimizar el posicionamiento. En los informes de resultados se relacionan las posiciones obtenidas al gestionar la aplicación con búsquedas en los motores seleccionados.

2.2.3.9 SEO Administrator v3.11

El estudio de los rankings es el principal servicio que presta la herramienta *SEO Administrator v3.11*[14] (http://www.seoadministrator.com/). Entre las funcionalidades básicas que incluye están la selección de palabras clave, el análisis de páginas web y el registro de las visitas de los usuarios. La interfaz resulta muy intuitiva permitiendo un cómodo manejo de la aplicación.

Los informes de resultados permiten analizar el grado de optimización que ha conseguido un sitio web atendiendo a las posiciones asignadas por una amplia lista de buscadores web. La herramienta también puede mostrar los rankings de los sitios en cada uno de los centros de datos de Google que se encargan de personalizar los resultados.

Además permite explorar los Snippets (almacenes virtuales de código público) con los que trabajan los buscadores web y a los que pueden acceder los usuarios para hacer

[14] Copyright 2002-2002 © Flamingo Soft

modificaciones. Esta herramienta también realiza un análisis de los enlaces de las páginas web para cada motor de búsqueda incluyendo en los resultados, si se desea, el valor del PageRank de Google.

La herramienta está preparada para registrar las visitas y extraer información mediante estadísticas de los hábitos y comportamientos de los usuarios. Los datos que se registran son concernientes a los accesos a las páginas (introducción directa de la URL o acceso por un resultado de un buscador), las palabras clave, los países de origen de las visitas y las navegaciones realizadas, incluidas las páginas de entrada y salida de las visitas. Los buscadores que intervienen es estos estudios son: Google, Yahoo Search y MSN. Otras visitas que se registran son las de los robots de los motores de búsqueda que acceden al sitio para indexarlo y hacerlo visible en sus resultados.

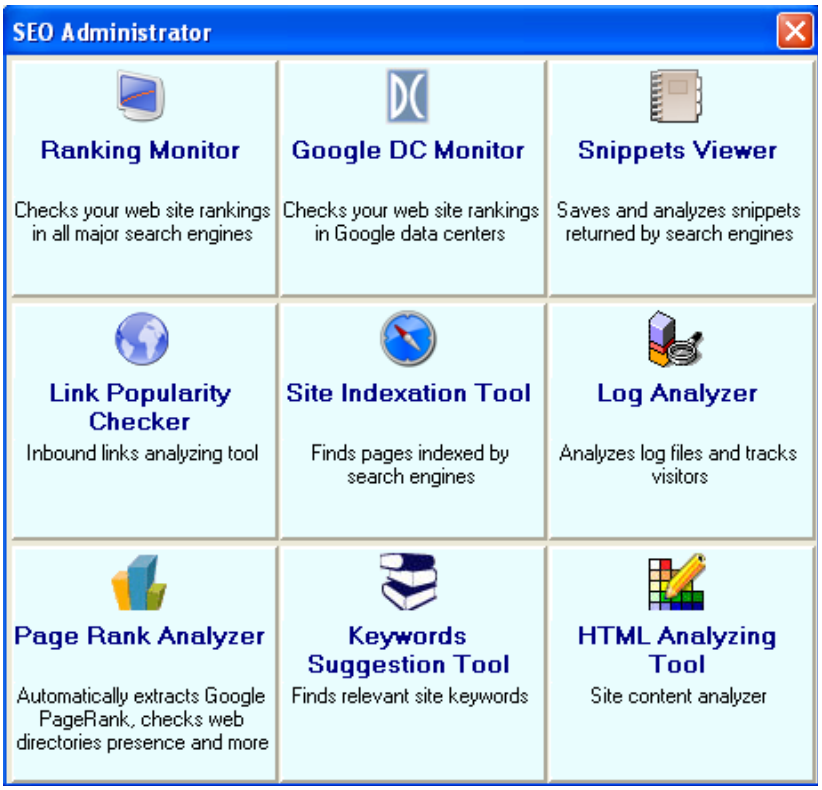


Figura II-23: Interfaz de Agent Web Ranking

Para el análisis de palabras clave la herramienta ofrece valores estadísticos comparativos de los sitios de la competencia. Las variables de estudio concretas se muestran en Tabla II-12.

Por último la aplicación incorpora un analizador de código HTML de páginas web que desglosa la información de la página por apartados de código.

VARIABLES DE ANÁLISIS DE LAS PALABRAS CLAVE DE LA COMPETENCIA	
Average PageRank	Media del PageRank de los competidores
Total number of page matching keywords	Número total de sitios resultados de la búsqueda
Number of exacts matches	Número de resultados exactos
Average Number of Inbounds links	Media del número de enlaces internos de la página web

Tabla II-12: Variables de análisis de las palabras clave de la competencia

2.2.3.10 SEO Elite

La aplicación *SEO Elite*[15] (http://www.seoelite.com/) destaca por su política de optimización basada en el incremento del número de enlaces externos mediante asociaciones con otras páginas. También analiza los rankings de las páginas web para las palabras claves seleccionadas considerando el número de páginas indexadas por los motores de búsqueda.

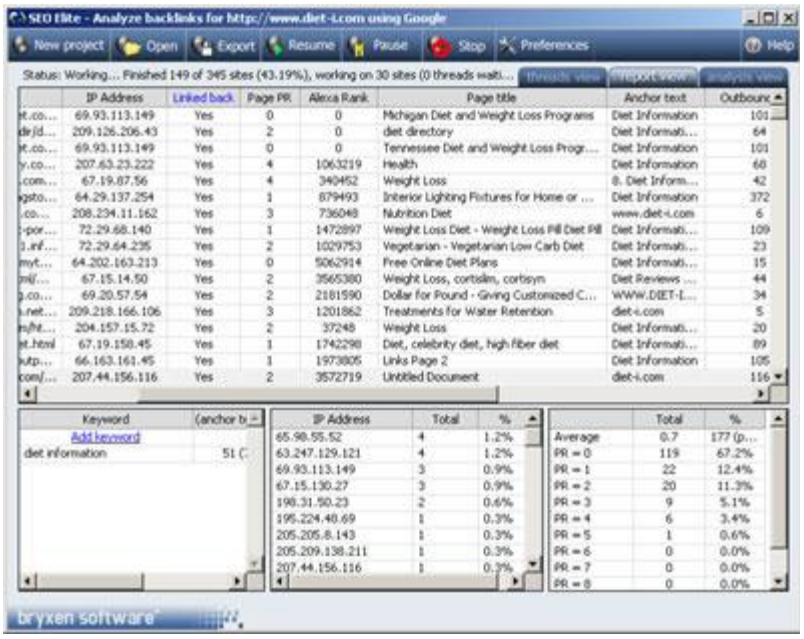


Figura II-24: Interfaz SEO Elite

En el estudio de los enlaces externos que apuntan a nuestro sitio web se puede hacer un análisis selectivo de los sitios que contengan determinadas palabras clave y elegir entre los buscadores Google, Yahoo, Altavista, All the Web o MSN.

Para la selección de socios con los que intercambiar enlaces se introducen los términos representativos de la temática deseada y el número mínimo de PageRank que deben tener

[15] Copyright 2005-2006 © Bryxen Software - A Search Engine Optimization Software

esos sitios web. Las secciones donde se buscan las palabras clave en los sitos web son el texto, el titulo y los enlaces. La herramienta utiliza una API de Google para que este buscador le permita realizar búsquedas automatizadas, práctica a la que habitualmente se opone Google.

Una vez que se ha alcanzado algún acuerdo se verifican los enlaces y se comprueba periódicamente el estatus de las páginas de los socios ya que puede disminuir la calidad de los enlaces.

2.2.3.11 1st Position Version 2.5.2.1

La herramienta SEO 1st Position Version 2.5.2.1[16] (http://www.1stposition.net/) centra sus rutinas de promoción, principalmente Altavista y Google, en el análisis de palabras clave. Seleccionando los términos adecuados para la optimización los autores esperan un aumento del tráfico en la página web objetivo y alcanzar una posición en el Top 10 de los rankings de los motores de búsqueda.

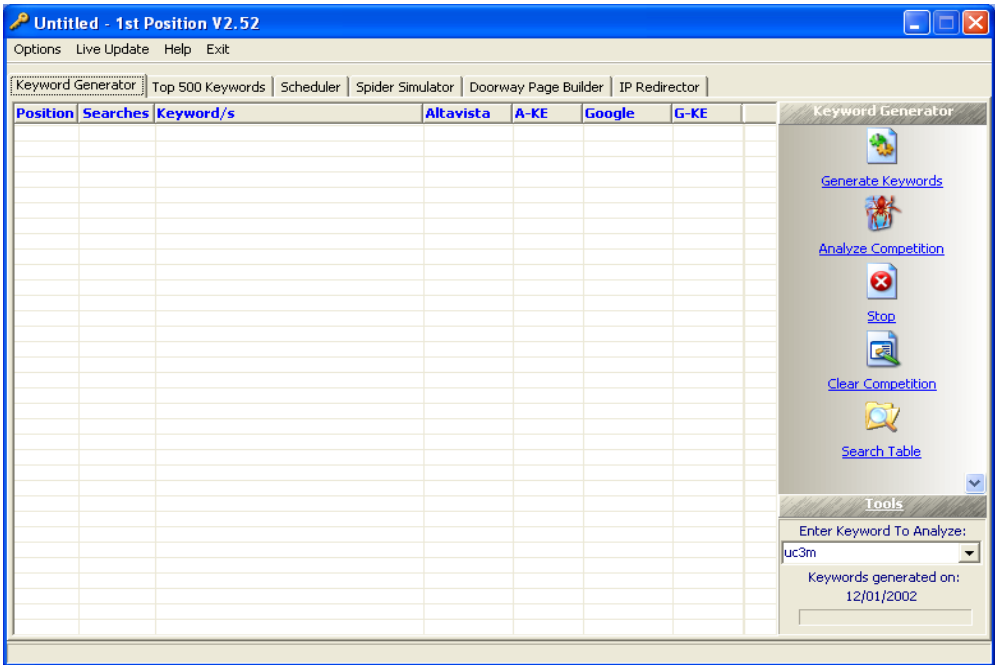


Figura II-25: Interfaz 1st Position

El análisis de términos óptimos genera alternativas a las palabras seleccionadas inicialmente por los diseñadores indicando lo adecuado de cada alternativa. Se muestran las palabras generadas y los datos de la competencia en los motores de búsqueda de Google y Altavista. Entre la información que aporta la aplicación se incluye:

[16] Copyright 2000-2002, TSRh Team

- El ranking de popularidad de las palabras clave mostradas.
- La media de búsquedas realizadas con esos términos.
- Las 100 palabras clave más populares relacionadas con las propuestas.
- El número de páginas resultantes en las búsquedas realizadas con esos términos.
- La efectividad de la palabra clave (relación del número de búsquedas y las páginas devueltas en esas consultas).

Otra opción para identificar las mejores palabras sin tener que proponer términos candidatos consiste en buscar en la clasificación de las mejores palabras claves de la red. Esta clasificación se basa en estudios mensuales y se actualiza online. Si se encuentra alguna que describa adecuadamente el mensaje de la página web a promocionar será la mejor opción posible.

La herramienta dispone de un simulador de spider que le permite visitar las páginas que la competencia muestra realmente a los robots de los motores de búsqueda. De esta forman se garantizan análisis correctos de la páginas de la competencia aún en el caso de que utilicen técnicas de engaño que muestran versiones diferentes de su web a robots y usuarios.

Para terminar, aunque ya se advirtió que el uso de *Doorway* puede ser penalizado si es detectado por buscadores, esta herramienta ofrece un constructor de este tipo de páginas estáticas optimizadas.

2.2.3.12 Good Keywords Gold

Good Keywords Gold[17] (http://www.goodkeywords.com/) es una herramienta SEO gratuita, con funcionalidades competentes para la selección de términos adecuados de búsqueda y optimización, además del análisis de popularidad de un sitio web.

[17] Copyright 1999-2006 Softnik technologies.

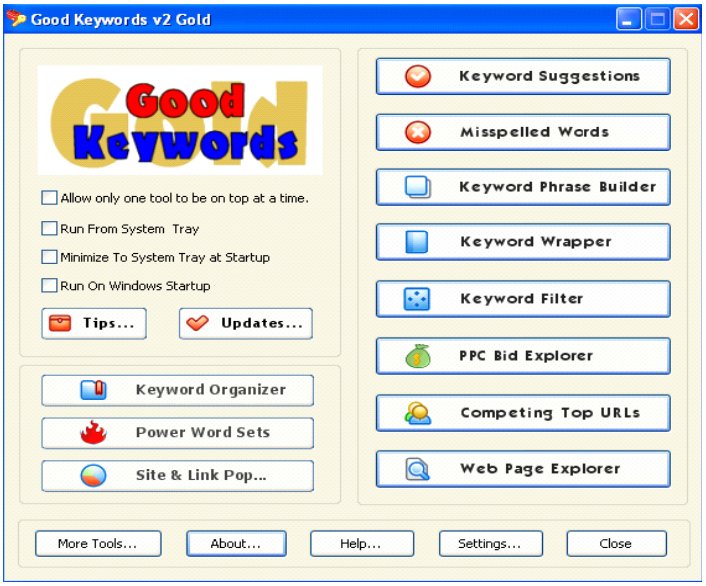


Figura II-26: Interfaz Good Keywords Gold

En la fase de construcción del conjunto de palabras clave, la aplicación analiza el interés de los usuarios en páginas relacionadas con los términos candidatos, examinando las búsquedas del último mes. Además sugiere palabras relacionadas con términos propuestos por los diseñadores, ya que pueden describir igualmente la página a promocionar y obtener mejor valoración que la seleccionada. La aplicación también examina las páginas situadas en el Top 10 determinando la competencia existente para cada palabra clave.

Esta herramienta rentabiliza los errores más comunes que cometen los usuarios al introducir en los buscadores los términos de consulta. Estos errores implican pequeñas variaciones en la escritura de las palabras. Analizando las variaciones de las palabras clave de un sitio web se pueden encontrar patrones de equivocaciones frecuentes. La aplicación propone seleccionar estos términos incorrectos como palabra clave, de esta forma una web optimizada obtendrá un buen posicionamiento incluso cuando las consultas contengan equivocaciones.

En cuanto a temas de popularidad, la herramienta puede promocionar web con técnicas de *Pay Per Clic*, analizar el tráfico de los sitios web con la herramienta Alexa e informar del número de enlaces en los buscadores Google, MSN y Altavista.

2.2.3.13 The Batch HTML Tidy Utility

La herramienta *The Batch HTML Tidy Utility*[18] (www.trellian.com/webtidy) forma parte del conjunto de aplicaciones desarrolladas por Trellian para la gestión de páginas web.

Herramienta SEO encargada de analizar el código HTML de las páginas web y de corregir sus posibles errores. El resultado es un código HTML de presentación profesional que permite una correcta lectura por exploradores o spiders. Por tanto, se garantiza una correcta indexación de las páginas web en los índices de los motores de búsqueda. Las opciones de la aplicación se muestran en la siguiente Figura II-27.

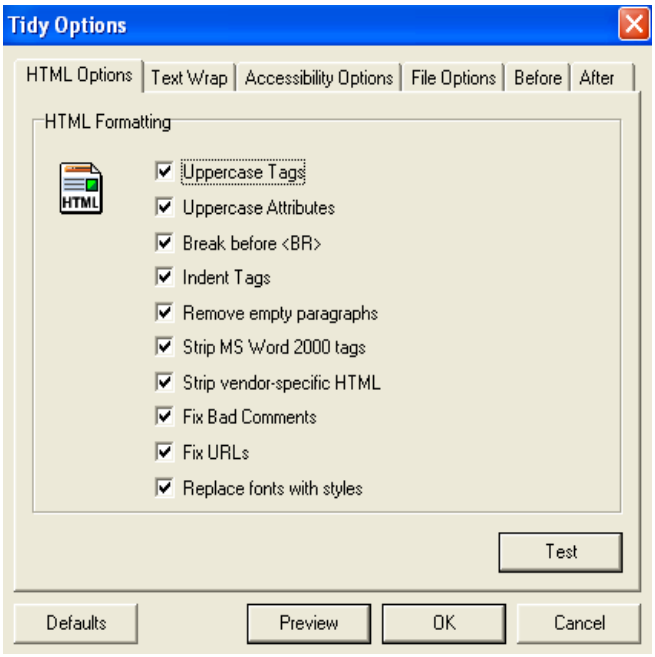


Figura II-27: Interfaz The batch HTML tidy utility

Los beneficios obtenidos al utilizar la herramienta *The Batch HTML Tidy Utility*, según los comentarios de sus autores son:

- Crea un código limpio, funcional y amigable para los motores de búsqueda.
- Permite una carga más rápida de la página web mejorando la accesibilidad.
- Soluciona los problemas de mal código existente.
- Elimina todas aquellas etiquetas innecesarias.
- Ayuda a mejorar la productividad y el tiempo invertido en la escritura de las páginas web.

[18] Copyright 2002-2003, Trellian Ltd.

- Ayuda a crear páginas como los webmasters.

Los campos que controla en el análisis del código HTML se muestran en la Tabla II-13.

CAMPOS DE CONTROL	
Uppercase Tags	Etiquetas en letras mayúsculas.
Uppercase atributtes	Atributos en letras mayúsculas.
Break before 	Terminación antes de .
Indent tags	Etiquetas con sangría.
Remove empty paragraphs	Eliminar párrafos vacíos.
Strip MS Word 2000 Tags	Etiquetas de MS word 2000
Strip vendor – specific HTML	Html especifico del fabricante.
Fix Bad comments	Corregir los malos comentarios.
Fix URL´s	Corregir URL´s.
Replace Font with styles	Reemplazar Fuentes con estilos.

Tabla II-13: Campos de control del HTML Validador de la herramienta The Batch HTML Tidy Utility

No obstante, aunque ayuda escribir HTML validado, no siempre es fundamental. Los ingenieros de Google, como Matt Cutts, han comentado que el algoritmo de posicionamiento de Google no tiene en cuenta, en absoluto, la corrección del código HTML de las páginas web (López, 2009).

2.2.3.14 Google Trends

Google Trends[19] (http://www.google.com/trends) permite analizar las búsquedas realizadas por los usuarios a lo largo del tiempo (hasta principios del 2004) apreciando cambios, modas y tendencias.

Su mayor utilidad, en cuanto a posicionamiento, se muestra a la hora de seleccionar palabras claves representativas de nichos de mercado concretos. Para ello se introducen los términos candidatos y se obtiene gráficas comparativas de cada palabra.

[19] ©2006 Google

Figura II-28: Interfaz Google Trends

Figura II-29: Resultados gráficos de Google Trends

La interpretación de las gráficas puede aportar información sobre las repercusiones que pueden tener ciertos eventos, en las preferencias de los usuarios por los términos de búsqueda. Se completan los resultados con la información de los lugares desde los que se han realizado las búsquedas de las palabras estudiadas. Se muestran 10 sitios ordenados y organizados según:

- Las ciudades origen de las búsquedas.
- Los países origen.
- El lenguaje del origen de las búsquedas.

2.2.3.15 Google Rankings

Google Rankings[20] (www.googlerankings.com) es una herramienta SEO cuyo propósito es la comprobación de las posiciones que ocupa una web, para determinadas palabras clave, en los diferentes rankings de los motores de búsqueda (Google, Yahoo Search o MSN).

Figura II-30: Interfaz de Google Ranking

Los parámetros de entrada para la aplicación son:

- Palabra o frase de palabras claves con los que obtener el ranking.
- Dominio o URL del sitio estudiado para hallar la posición.
- Opción de hallar ranking con Yahoo y MSN incluidos.

Los resultados se dividen por motores de búsqueda:

- Google:
  - Posición de la página.
  - PageRank de la página.
  - Texto de la página mostrado en la búsqueda.
  - Número de enlaces externos.

[20] Google™

  - Tiempo utilizado para la obtención de los datos.

- Yahoo: Posición, página y tiempo.
- MSN: Posición, página y tiempo.

2.2.3.16 Google Suggest

Google Suggest[21] (www.labs.google.com/suggest) es una herramienta online de Google encargada de proponer términos candidatos a palabras claves y consultando por ellos, mostrar los resultados. También puede sugerir alternativas a palabras clave incluidas de forma preliminar. La herramienta muestra una interfaz y funcionamiento similares al motor de búsqueda Google.

Figura II-31: Interfaz de Google Suggest

Los resultados ayudarán a los diseñadores a seleccionar aquellos términos que siendo buenos descriptores ofrezcan menos competencia.

2.2.3.17 Google Analytics

Google Analytics[22] (http://www.google.com/analytics/) es una herramienta online ofrecida por Google de modo gratuito cuya misión fundamental es el análisis de las consultas realizadas por los usuarios cuando acceden a una determinada página web. Previamente se ha de insertar unas líneas de código de rastreo para Google en el código HTML de la página en estudio. Así se registran las visitas y se generarán los informes.

[21] ©2006 Google
[22] ©2006 Google

Figura II-32: Interfaz de Google Analytics

Los análisis proporcionan, entre otros, los siguientes datos:

- Palabras que más utilizan los usuarios que acceden a la web en estudio. Estos datos permiten seleccionar como palabras clave aquellas que más utilicen los usuarios para encontrar la página web en la red.
- Número de visitantes que proceden de un motor de búsqueda y número de usuarios que introducen directamente la dirección URL. Si la mayoría de usuarios acceden desde buscadores se puede plantear campañas de promoción en los motores de búsqueda.
- Origen de las búsquedas. La información sobre el motor de búsqueda utilizado y el lugar del mundo desde donde se accede a la página puede orientar o particularizar la oferta de la página.
- Número de usuarios que vuelven a acceder al sitio. Es un indicador del atractivo de la página.
- Páginas vistas en el dominio por los usuarios.
- Rastreo del retorno de las inversiones (ROI) realizadas en promoción web.

2.2.3.18 Google Sitemaps

Google Sitemaps[23] (http://www.google.com/webmasters/sitemaps/) es una herramienta de Google a modo de medio de comunicación entre los webmasters y este motor de búsqueda. Los sitios web proporcionan a Google información que es aprovechada para indexarlos de forma más eficaz, del otro lado, Google evalúa las páginas web e indica los problema que han surgido al rastrear el sitio. De esta forma los webmasters pueden corregir los errores que repercuten en una inadecuada indexación.

Se puede añadir un Sitemap en una cuenta de Google para proporcionarles información sobre las páginas de un sitio y de este modo ayudarles a rastrearlas de forma más eficaz. O bien, se puede simplemente añadir un sitio a una cuenta Google para visualizar la información disponible sobre el sitio en cuestión.

Los webmasters pueden incorporar información adicional sobre cada una de las URL: última actualización, frecuencia de las modificaciones, grado de importancia en comparación con las demás URL del mismo sitio. Estos datos permiten rastreos más inteligentes de los sitios.

Algunas de las estadísticas proporcionadas por Google sobre un sitio web analizado son:

- Las páginas indexadas del sitio web.
- Las páginas que referencian la URL del sitio web.
- Las páginas que enlazan con el sitio web.
- Caché actual del sitio web.
- Información disponible sobre el sitio web.
- Páginas similares al sitio web.

2.2.3.19 Traffic Rankings de Alexa

Alexa[24] (http://www.alexa.com/) es un sistema de evaluación utilizado y aceptado como parámetro de referencia en el ranking de popularidad por los grandes sitios y las grandes empresas en internet. Maneja información sobre sitios web relacionados, estadísticas de visitas, valoración de los usuarios, propietarios, fecha de creación y además realiza

[23] ©2006 Google
[24] ©1996-2006, Alexa Internet, ENC.

comparativas de tráfico con otros sitios analizando semanalmente las tendencias en visitas y páginas vistas.

El ranking de Alexa está basado en las visitas de los internautas que tienen instalada su barra (más de 10 millones en todo el mundo) en períodos de tres meses. La posición que ocupa un sitio en el ranking mundial es una combinación del alcance y páginas vistas obtenidas, definiéndose estos parámetros como:

- Alcance (*reach*): número de usuarios (direcciones IP) que visitan un sitio en un día dado.
- Páginas visitadas (*page views*): cantidad de páginas visitadas por las urls diferentes que visitan un sitio. En distintos días, la misma url se cuenta como diferente.

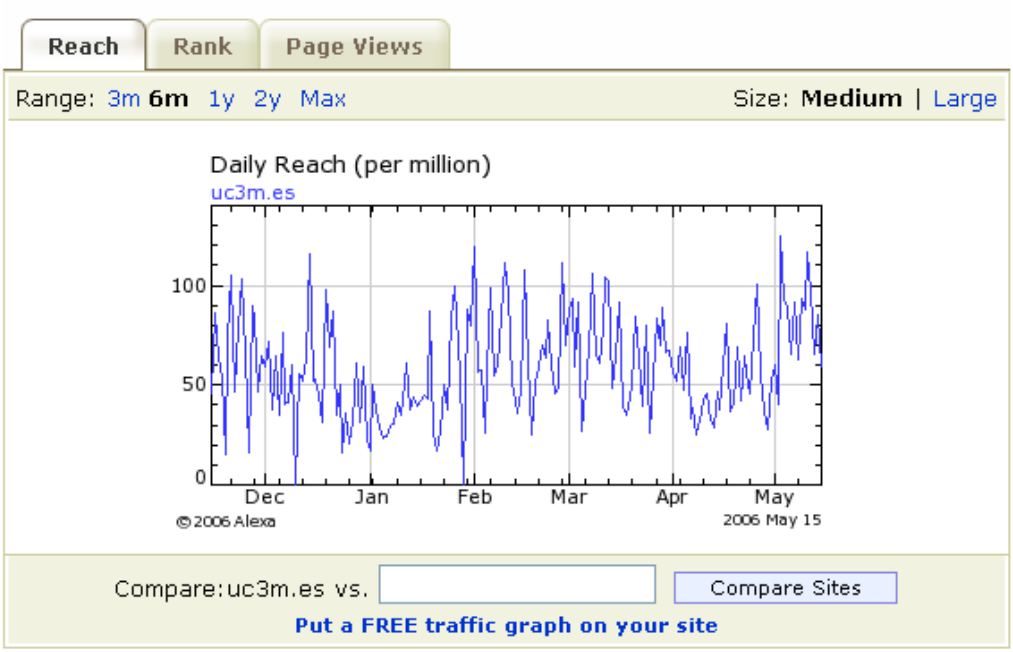


Figura II-33: Ejemplo gráfico del Tráfico en Alexa

En la interpretación de los datos que proporciona esta herramienta se deben tener en cuenta los siguientes sesgos:

1. Funcionamiento limitado a las webs de nivel superior del tipo www.dominio.com.
2. Sólo funciona con el navegador *Internet Explorer* y el sistema operativo *Windows*.
3. Se desactiva en las páginas seguras (*https:*) de los sitios.
4. Los sitios con una posición por encima del puesto 100.000 no son fiables (con menos de 1.000 visitantes mensuales), ya que la cantidad de datos obtenida no es estadísticamente significativa.

5. Los factores culturales y la lengua (frecuentemente el inglés) influyen en la adopción de su software.

2.2.3.20 ToolbarBrowser

ToolbarBrowser[25] (http://www.toolbarbrowser.com/support.htm) es una herramienta SEO con interfaz tipo *toolbar*. La barra de herramientas se descarga y se añade al explorador para poder acceder a sus funcionalidades directamente al visitar páginas web.

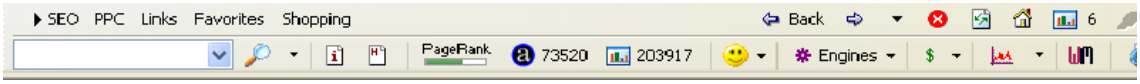


Figura II-34: Barra de herramientas de Toolbar Browser

De entre las múltiples funcionalidades de la herramienta se enumeran y describen las que tienen más interés para el posicionamiento web (módulo SEO visible en la barra).

FUNCIONALIDADES SEO DE LA HERRAMIENTA TOOLBAR BROWSER	
	Simulador de Spider que muestra como ven el sitio los buscadores
	Cabecera del servidor
PageRank	Ranking de Google PageRank * popularidad de los enlaces
73520	Ranking de Alexa * popularidad del trafico de la página
203917	Ranking de las páginas indizadas
	Enlaces a foros temáticos de optimizaron
Engines	Enlaces a una serie de motores de búsqueda
$	Enlaces los sitios con cuentas de *Pay Per clic*
	Detalles del trafico de la página según Alexa.
	Mostrar Versiones anteriores de la página con *Way Back Machine*
Keyw »	Estudio de palabras clave con la herramienta *Keyword Discovery*
	Opciones de configuración y ayuda de la barra de herramientas

Tabla II-14: Funcionalidades SEO de Toolbar Browser

[25] Copyright 2004, 2005 ToolbarBrowser.com.

2.2.3.21 SEOpen Toolbar

La herramienta SEOpen Toolbar[26] (http://www.seopen.com/) es una aplicación cuya interfaz es una barra de herramientas diseñada para Firefox.

Figura II-35: Barra de herramientas de SEOpen Toolbar

Dispone de diversas funcionalidades comunes y particulares para cada motor de búsqueda (Google, Yahoo Search, MSN y Alexa.). La Mass Check da acceso a las funcionalidades compartidas por los cuatro buscadores:

- Número de enlaces de la página web (*Backlinks*).
- Sitios indexados por el buscador (*Indexed*).
- Enlaces que contiene el dominio (*Domainlink*).

Sólo Google y Alexa tienen asociadas funcionalidades específicas en esta herramienta. Se muestran en la Tabla II-15 únicamente las relativas a Google, ya que las de Alexa coinciden con las descritas en su propia herramienta:

FUNCIONALIDADES ESPECIFICAS DE GOOGLE	
Cache	Ultima versión de la copia en cache guardada
Relamed	Sitios relacionados con la página web
Traslate	Traducción de la página web
Check PR	Chequeo del PageRank de Google

Tabla II-15: Funcionalidades SEOpen Toolbar para Google

En la opción *Other Tools*, se tiene acceso a otras herramientas de interés. Estas se enumeran en la siguiente tabla junto a su cometido (Tabla II-16):

[26] SEOpen.com © 2005 Damian Smith

HERRAMIENTAS	FINALIDAD
Check DMOZ	Chequeo de la inclusión de la página en el directorio DMOZ
Link analizer	Análisis de los enlaces en la herramienta de SEO Tools ya estudiada
Keyword density	Análisis de la densidad de las palabras clave en la herramienta SEO Tools
Page size	Muestra el tamaño de la página
HTML Validator	Chequeo del código HTML
Header Server	Cabeceras del servidor
Robots.txt	Fichero de permisos para los spiders
Archives	Acceso aversiones anteriores de la página ofrecidas por *Way Back Machine*
WHOIS	Describe las páginas web según el directorio DMOZ y Alexa

Tabla II-16: Herramientas de interés incluidas en SEOpen

2.3 Técnicas de aprendizaje automático

2.3.1 Introducción

Las técnicas de aprendizaje automático son mecanismos para la obtención de patrones a partir de recopilaciones de datos. Estos datos se suelen presentar a los algoritmos como ejemplos de aprendizaje.

Dependiendo de que estos ejemplos estén ya clasificados a priori o no, se pueden dividir estas técnicas en dos categorías: técnicas supervisadas y técnicas no supervisadas (Weiss y Indurkhya, 1998).

En las técnicas supervisadas se conoce la clase a la que pertenece cada ejemplo de aprendizaje. A su vez estas técnicas se pueden clasificar en dos grupos: técnicas predictivas y técnicas de clasificación (Weiss y Kulikowski, 1991).

2.3.2 Técnicas de inducción reglas

La inducción de reglas es una técnica que recibe la información como un conjunto de casos (como ejemplos de aprendizaje). Estos ejemplos se representan por un conjunto de atributos común que incluye el atributo de clase. Los valores de estos atributos distinguen unos casos de otros.

Las técnicas de inducción de reglas a partir de los datos de entrada generan un árbol de decisión o un conjunto de reglas que proporcionará la clasificación de los nuevos ejemplos (Hong et al., 1986; Clark y Niblett, 1989).

Existen dos estrategias principales para conseguir la inducción de reglas:

1. Generación de un árbol de decisión y posterior extracción de sus reglas (Quinlan, 1993).
2. Aplicación de una estrategia de *covering* que genere reglas que cubran todos los ejemplos de una única clase y tras eliminar los ejemplos cubiertos proseguir con otra clase.

Un ejemplo de aplicación de la primera estrategia es el sistema C 4.5 (Quinlan, 1993), una extensión del ID3 (Quinlan, 1986), que puede generar reglas previa generación de un árbol de decisión.

Otros algoritmos como el PRISM (Cendrowska, 1987) se basan únicamente en una estrategia de *covering*, mientras que algoritmos como el PART (Frank y Witten, 1998) combinan ambas estrategias.

Entre las ventajas que presentan estas técnicas destacan:

- Robustez frente al ruido (debidos a errores, omisiones o insuficiencia de datos).
- Identificación de atributos irrelevantes.
- Detección de la ausencia de atributos discriminantes y de vacios de conocimiento.
- Extracción de reglas fáciles de entender y de gran expresividad.
- Posibilidad de reprocesar las reglas mediante el conocimiento de expertos, interpretando, modificando o aceptando reglas (Major y Mangano, 1995).

2.3.3 Conjuntos de Clasificadores

Los clasificadores son modelos que a partir de los valores de las características que representan a un elemento proporcionan la categoría a la que pertenece.

El objetivo de los conjuntos de clasificadores es mejorar la precisión que obtendrían de forma individual cada uno de los clasificadores pertenecientes al conjunto. Las decisiones que toma cada clasificador frente a un nuevo ejemplo son combinadas obteniéndose una única decisión final (Dietterich, 1997).

Entre las técnicas de construcción de conjuntos de clasificadores más comunes destacan: los clasificadores homogéneos y los clasificadores heterogéneos.

2.3.3.1 Clasificadores homogéneos

Los clasificadores homogéneos son generados a partir del mismo algoritmo de aprendizaje (Dietterich, 2000). Los principales métodos de construcción de clasificadores homogéneos son *Bagging* (Breiman, 1996) y *Boosting* (Schapire, 1990).

Ambos métodos generan diferentes hipótesis mediante manipulación de los ejemplos de entrenamiento. El algoritmo base es entrenado con diferentes conjuntos de instancias construyéndose de esta forma cada uno de los clasificadores que forman parte del conjunto. La decisión sobre qué clasificadores proporcionan la predicción más oportuna, dependiendo de la instancia a clasificar, se realiza mediante un sistema de votos.

El éxito de estos métodos depende de manera directa de la diversidad de hipótesis generadas. Por tanto, funcionan mejor cuando el algoritmo base es un algoritmo de aprendizaje inestable, es decir, el algoritmo base construye modelos muy diferentes al presentarle distintos conjuntos de entrenamiento. Por ejemplo, los algoritmos de inducción de reglas son algoritmos de aprendizaje inestables y por consiguiente son adecuados como algoritmo base.

En concreto, *Bagging* (*Bootstrap Aggregation*) (Breiman, 1996) obtiene diferentes muestras a partir del conjunto de ejemplos de entrenamiento. Cada una de estas muestras contiene el 63,2% de instancias del conjunto original, en promedio, alcanzando el número de ejemplos total mediante repeticiones de instancias. El sistema de votos que unifica las predicciones de los clasificadores del conjunto está basado en la elección de la clase más votada por los clasificadores del conjunto.

En cambio *Boosting* (Schapire, 1990) en su versión más representativa (*AdaBoost*) (Freund y Schapire, 1995, 1996) construye los clasificadores de forma secuencial, centrándose especialmente en los ejemplos que han sido clasificados de forma errónea por el último clasificador generado. Cada ejemplo de aprendizaje recibe la asignación de un peso en función de la dificultad que presenta al intentar clasificarlo acertadamente. Estas ponderaciones son actualizadas conforme evoluciona el algoritmo. Se utiliza en este algoritmo una estrategia de voto ponderado con el fin de discernir la decisión final. El peso de cada clasificador en la votación está en concordancia con la precisión que obtiene

sobre el conjunto de entrenamiento utilizado en su generación. Se tiene en cuenta en este proceso los pesos de las instancias.

2.3.3.2 Clasificadores heterogéneos

Los clasificadores heterogéneos se generan a partir de distintos algoritmos de aprendizaje. El método más utilizado en la construcción de este tipo de clasificadores es *Stacking* (*Stacked Generalization*) (Wolpert, 1992). Este método combina las predicciones de un conjunto de clasificadores de diferente tipología mediante otro algoritmo de aprendizaje. Es decir, se implanta un clasificador en un nivel superior cuyo cometido es aprender a combinar los resultados del resto de los clasificadores.

El problema más común en la aplicación de *Stacking* es la elección de los algoritmos de aprendizaje para obtener los clasificadores base y el meta-clasificador.

2.3.4 Métodos de selección de atributos

En los problemas de clasificación automática los individuos se identifican por los valores que toman en el conjunto de atributos que los representa. Estos atributos se pueden filtrar atendiendo a su valía en la clasificación mediante métodos de selección. Estos métodos reducen el número de atributos del conjunto inicial (Molina y Belanche, 2002), o los ordenan por su importancia.

La selección de atributos combina dos algoritmos, uno de búsqueda, con la estimación de la utilidad del atributo, y otro de evaluación.

En los casos en que los algoritmos de selección evalúan los atributos apoyándose en un esquema de aprendizaje (Hall y Holmes, 2002), se les denominan *Wrappers* (Morales, 2007), en caso contrario se les llama filtros.

Desde otro punto de vista también se puede distinguir a los métodos de selección por evaluar a los atributos de forma independiente, o evaluar subconjuntos de atributos. Estos últimos a su vez admiten otras subclasificaciones atendiendo a las técnicas de búsqueda de subconjuntos de atributos (Hall y Holmes, 2002).

Los siguientes algoritmos son ejemplos de filtros, y por tanto no basan sus evaluaciones en métodos de clasificación:

- **CfsSubsetEval:** Evalúa los subconjuntos de atributos por la calidad de las predicciones individuales de cada variable. Los subconjuntos de atributos muy

correlacionados con la clase y con bajo nivel de intercorrelación son los mejor valorados (Hall, 1998).

- **ConsistencySubsetEval:** Evalúa un subconjunto de atributos por el nivel de consistencia en los valores de la clase al proyectar las instancias de entrenamiento sobre el subconjunto de atributos (Liu y Setiono, 1996).

Los métodos *Wrappers*, al utilizar un algoritmo de aprendizaje para medir lo deseable de un subconjunto de atributos, son apropiados en la generación de clasificadores más precisos por eliminación de atributos redundantes o dependientes de los demás. Los siguientes métodos de selección son ejemplos de *Wrappers*:

- **ClassifierSubsetEval:** Evalúa los subconjuntos de atributos por medio de un clasificador, bien con los datos de entrenamiento, o bien con un conjunto de test.
- **WrapperSubsetEval:** En este algoritmo se emplea validación cruzada en las estimaciones de calidad de los subconjuntos de atributos (Kohavi y John, 1997).

Los evaluadores de atributos individuales, al no eliminar atributos redundantes, son adecuados para proporcionar listas ordenadas de todos los atributos según su calidad, con independencia de los demás. Ejemplos destacados de este tipo de evaluadores de atributos son:

- **ChiSquaredAttributeEval:** Obtiene el nivel de correlación entre la clase y cada uno de los atributo calculando el valor estadístico Chi-cuadrado (Freund et al., 2000).
- **GainRatioAttributeEval:** Evalúa los atributos examinando su razón de beneficio con respecto a la clase.
- **InfoGainAttributeEval:** Después de discretizar los atributos numéricos, se calcula la ganancia de información de cada atributo con respecto a la clase (Lorenzo, 2002).
- **OneRAttributeEval:** También discretiza los atributos numéricos. Evalúa los atributos con el clasificador OneR (Holte, 1993).

2.4 Trabajos previos asociados

Se han estudiado trabajos de investigación que tienen relación con la predicción de ranking web, aunque ni el planteamiento del problema ni las metodologías empleadas coinciden con el planteamiento de esta investigación. Varios de estos trabajos señalan la necesidad de hacer predicciones del PageRank aludiendo a sus tardías actualizaciones. Según sus autores el constante incremento del tamaño de la Web alarga en el tiempo las exploraciones necesarias para su cálculo.

En los trabajos de Chien et al. (2003) se presenta un sistema de estimación del PageRank en base a las modificaciones de un conjunto de enlaces. Su método consiste en restringirse al subgrafo formado por esos vínculos y su vecindario más cercano, conectándolo con un nodo más que representa toda la Web. El cálculo del PageRank se realiza de forma rápida sobre este grafo reducido.

En la investigación de Yang et al. (2005) se propone un método predictivo con una finalidad diferente. En este caso no se intenta predecir el PageRank, la idea es demostrar que sus ordenaciones no son precisas y dar una alternativa. Las características de las direcciones URL se han utilizado para hacer predicciones del PageRank mediante regresión lineal (Kan, 2005). También en este trabajo se aplica el mismo enfoque para determinar la pertinencia de una página web a una consulta. Es decir, se indica si es un resultado adecuado para la consulta, pero no el grado de pertinencia en comparación con otros documentos relevantes.

Las cadenas de Markov se han empleado en el trabajo de Vazirgiannis et al. (2008) para la predicción de rankings de páginas web, para ello necesita de históricos de ordenaciones previas como medio de obtención de las tendencias en la Web. Con la misma filosofía de predecir futuros rankings a partir de ordenaciones anteriores en (Zacharouli et al., 2009) se emplean técnicas de regresión lineal, clustering y análisis de las componentes principales sobre los valores de PageRank y tf-idf de las páginas web. Los valores de tf-idf se calculan sobre términos de consulta escogidos al azar.

Las técnicas de aprendizaje automático se han aplicado en la construcción de modelos de ordenación.

En particular, los algoritmos genéticos tienen una gran trayectoria investigadora en el campo de la recuperación de información. En 1998, Gordon presenta un algoritmo genético que tomando la función de Jaccard como fitness y a las palabras clave como

genes obtiene el conjunto de palabras que mejor describen a los documentos (Gordon, 1988). Es decir, se trata de un sistema de filtrado, si bien las relaciones semánticas no están tratadas en el trabajo. Posteriormente, Morgan y Kilgour (1996) estudian una población de consultas que mutan y se van recombinando. El proceso facilita la incorporación de nuevos términos como sinónimos utilizando un diccionario. La aplicación del resultado es una recuperación personalizada.

Curiosamente con los genéticos se han observado conclusiones similares a las de Zipf, como se ve en los trabajos de Cummins y O'Riordan (2005). Estos autores proponen, como en otros trabajos, esquemas de pesos para aplicar algoritmos genéticos. De este modo se consolida la idea de que los términos más resolutivos son aquellos de frecuencia de aparición intermedia ya que son los que obtienen las asignaciones de los pesos más altos.

En el trabajo de Trotman (2005) se hace énfasis en la diferente influencia que tiene la aparición de un término en las distintas estructuras que componen un documento. En concreto, se trabaja con documentos XML. Las ponderaciones de las estructuras conforman los cromosomas que son manipulados por un algoritmo genético para optimizar el posicionamiento. El proceso es realizado y evaluado a partir de la colección TREC WSJ.

Otro tipo de algoritmos bioinspirados se basan en la emulación del sistema inmune (de Castro y Timmis, 2003). Su aplicación en problemas de optimización mejora los resultados obtenidos por algoritmos genéticos (Musilek et al., 2006).

En (Wang et al., 2010) se utiliza este enfoque para resolver el problema de metabúsqueda (ordenar una lista de resultados a partir de otras ordenaciones de los mismos).

Los experimentos se realizaron sobre las colecciones OHSUMED, TREC 2003 y 2004. Otras estrategias de aprendizaje aplicadas al mismo problema son las máquinas vectoriales (Joachims, 2002), (Cao et al., 2006) y las técnicas *Boosting* (Freund et al., 2003), (Xu y Li, 2007).

Otros trabajos sobre ranking comparan los resultados obtenidos para una misma consulta por diferentes motores de búsqueda web. La finalidad de esas investigaciones es averiguar el grado de similitud en el comportamiento de sus respectivos algoritmos de posicionamiento. Estos estudios demuestran, a lo a largo del tiempo, las diferencias existentes en los algoritmos de ordenación de resultados de los motores de búsqueda

(Ding y Marchionini, 1998), (Bharat y Broder, 1998), (Chignell, Gwizdka y Bodner, 1999), (Gordon y Pathak, 1999), (Nicholson, 2000), (Jux2.com, 2004), (Egghe y Rousseau, 2005) y (Barllan, 2005). En estos trabajos se evalúa el grado de solapamiento entre los resultados de diferentes buscadores web. Los experimentos también muestran diferencias en los algoritmos de indexación de términos y en las técnicas de búsqueda.

En un trabajo similar Dogpile.com (2007) se evalúan los principales motores de búsqueda Google, Yahoo, Windows Live ™ (antes MSN Search) y Ask ™ concluyendo diferencias entre sus resultados tanto a nivel global como en las primeras posiciones.

Parte de la investigación que se amplía en la presente tesis es fruto de varios trabajos que están en una línea diferente. El objetivo de ellos, (Moreno, 2005) y varios Proyectos de Fin de Carrera dirigidos por este autor, es la identificación de los factores de posicionamiento que afectan en las ordenaciones de los motores de búsqueda, la influencia que ejercen en estas ordenaciones y las correlaciones que hay entre ellos. En UC3M (2009), se utiliza la información de estos factores para crear modelos predictivos que determinen las posiciones que ocuparían las páginas web, tras un proceso de optimización, entre los resultados de una consulta.

Capítulo III: Desarrollo de la Investigación y Marco Experimental

Como se planteaba en la introducción de este trabajo, el objeto del presente estudio es establecer un método para la estimación de la relevancia documental de un documento para una consulta determinada respecto a los documentos de su competencia. A diferencia de otras propuestas, el objetivo es precisar la posición en el ranking que se le otorgará a una página web ante determinadas consultas.

Con este objetivo se han marcado tres fases de investigación, desarrollo e experimentación que se simplifican en la siguiente Figura III-1:

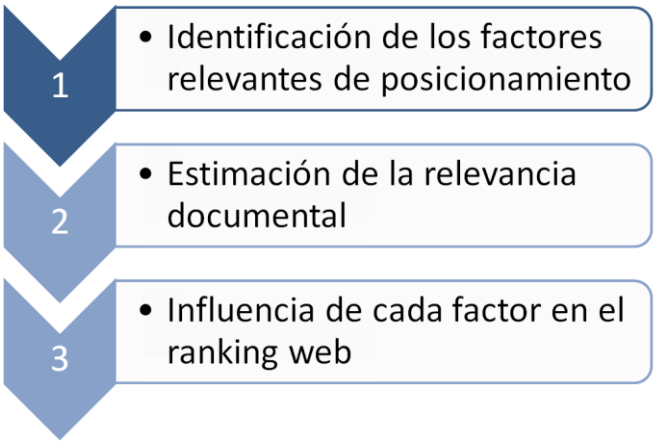


Figura III-1: Fases de la investigación

La fase 1 consiste en la identificación de los factores relevantes del posicionamiento web utilizados por las herramientas SEO (epígrafe 3.1). Para ello se analizará un conjunto de herramientas y se extraerán e identificarán los factores de posicionamiento que usan.

En la fase 2 sobre la estimación de la relevancia documental se estudia la relevancia que asignan los buscadores web (epígrafe 3.2) a las páginas web.

Por último en la fase 3 se investiga las posibilidades para determinar de forma automática la influencia de cada factor en los algoritmos de ordenación de los motores de búsqueda web (epígrafe 3.3).

Estas fases responden al planteamiento de definir la fase metodológica y aplicar una experimentación cuyos resultados serán evaluados para su validación.

3.1 Identificación de los factores de posicionamiento web en herramientas SEO

En este apartado el objetivo principal es la identificación o detección de los factores más relevantes para el posicionamiento web utilizados por las herramientas SEO. Para determinar cuáles son los factores de posicionamiento más determinantes, en este trabajo, se parte de la idea de que una variable de posicionamiento será más relevante cuanto mayor sea el número de herramientas que la utilizan.

Para realizar este primer estudio se llevan a cabo diferentes fases: una fase metodológica para identificar los factores de posicionamiento, una fase de análisis comparativo de herramientas, y por último una fase en la que se muestran los resultados y conclusiones.

3.1.1 Metodología para la identificación de factores de posicionamiento

La metodología seguida para la identificación de factores de posicionamiento se basa en un estudio comparativo de los factores y funcionalidades de las herramientas SEO (epígrafe 2.2.3). Se han seleccionado 39 herramientas SEO con identidad propia, y decenas de subherramientas integradas en ellas. De esas 39 herramientas propias, 21 son herramientas principales, es decir, no prestan sus servicios a partir de otras. La selección de estas herramientas responde a que son las aplicaciones SEO más utilizadas por los diseñadores en la optimización web y han sido estudiadas en el apartado del Estado del Arte.

Con el fin de identificar los factores y funcionalidades de posicionamiento web más relevantes en las herramientas SEO, se ha realizado un estudio comparativo que se plantea en las siguientes fases:

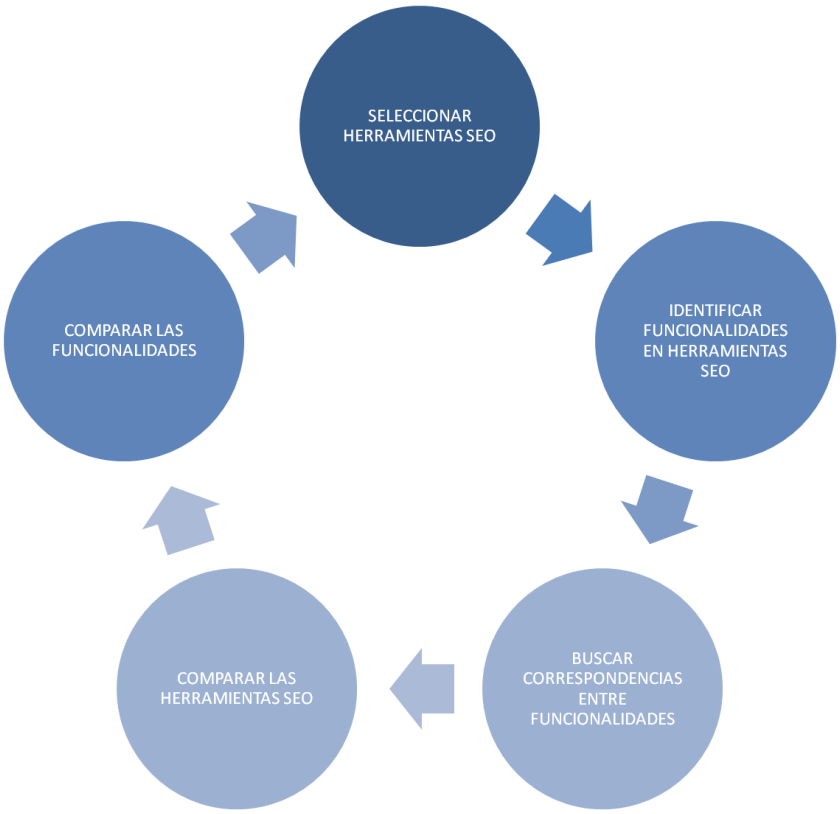


Figura III-2: Ciclo metodológico para la identificación de factores de posicionamiento

1. Identificar los factores de posicionamiento analizados por las herramientas SEO agrupándolos en funcionalidades para facilitar un análisis comparativo.
2. Buscar correspondencias entre las herramientas SEO para ver las funcionalidades que considera cada una.
3. Comparar las herramientas SEO y contabilizar el número de funcionalidades de posicionamiento que evalúan.
4. Comparar las funcionalidades SEO respecto al número de herramientas SEO que las analizan.

3.1.2 Análisis comparativo para determinar factores y funcionalidades de posicionamiento web

3.1.2.1 Identificación de las funcionalidades ofrecidas por las herramientas SEO

En el apartado 2.2.3 se han presentado las herramientas SEO más utilizadas por los diseñadores en la optimización web. Se ha realizado, para cada una de las aplicaciones

SEO, un análisis de los aspectos relacionados con el posicionamiento web. Se han identificado factores o variables SEO y se han sintetizado en 40 funcionalidades asociadas. La lista de funcionalidades de posicionamiento se detalla en la Tabla III-1, donde se indica la funcionalidad y una breve descripción.

FUNCIONALIDAD SEO		DESCRIPCIÓN
CHEQUEO DE ENLACES		Análisis de los enlaces por posibles problemas de conectividad y comunicación.
PRESENTACIÓN DE PÁGINAS	AUTOMÁTICA	Presentación de la información de la página a los motores de búsqueda y directorios sin visitar la página de presentación de cada uno.
	MANUAL	Acceso a las páginas de los motores de búsqueda para introducir la información.
PPC		Técnicas *Pay Per Click* en las que se patrocinan enlaces a cambio de un cobro por cada vez que un usuario pinche en el enlace correspondiente.
PAGO POR INCLUSIÓN		Presentación de las páginas a motores de búsqueda que cobran por ser incluidos en sus índices.
HTML VALIDADOR		Chequeo del código HTML de la pagina verificando que se haya construido de forma correcta.
CATEGORIZACIÓN		Métodos para otorgar una categoría a las páginas para ayudar a los motores de búsqueda a clasificar la pagina de forma correcta y a nuestra intención.
INFORMES		Se generan informes de resultados que se almacenan para futuros procesos.
BÚSQUEDAS PALABRAS CLAVE USUARIOS		Se trata el número de búsquedas realizadas con las palabras claves elegidas por los usuarios, en los motores de búsqueda en un determinado periodo, para verificar la popularidad de las palabras clave. También se sugieren palabras y frases que contienen las introducidas.
CONSTRUCTOR DE PÁGINAS		Herramienta de ayuda a la creación del código HTML de una página web.
BÚSQUEDAS EN SEARCH ENGINE		Búsquedas en motores de búsqueda.
MERCADO DE ENLACES		Gestión de intercambio de enlaces con otros sitios relacionados temáticamente.
RANKING PALABRAS CLAVE		Las palabras clave otorgan a cada página web un ranking que será el resultado de esta variable.
CREACIÓN PÁGINAS DOORWAY		Se ofrece la posibilidad de editar páginas doorway para la optimización de las páginas.
RASTREO DEL TRÁFICO		Rastreo del tráfico que recibe una página web: número de usuarios, páginas vistas, procedencia, etc.
GESTION DE DOMINIOS		Gestión de las licencias de los dominios de las páginas web.
COMPARACION PÁGINAS PRIMERAS POSICIONES		Se llevan a cabo análisis de las páginas de los primeros puestos para obtener patrones de optimización.

PROGRAMADOR	Se ofrece la posibilidad de automatizar las tareas con un programador de tareas.
POPULARIDAD ENLACES	Se mide el grado de popularidad de una página según el número y la procedencia de los enlaces que la apuntan.
CHEQUEO DMOZ DIRECTORY	Se chequea si la página ha sido incluida en el directorio de referencia DMOZ.
DENSIDAD DE PALABRAS CLAVE	Se analiza la frecuencia y densidad de aparición de las palabras clave en la página.
PÁGINAS INDEXADAS	Se analizan las páginas de un sitio Web indexadas por un determinado motor de búsqueda.
HISTORIAL DE CAMBIOS	Se registran los cambios de posición y otros factores de la página.
EDITOR HTML	Herramienta de edición de los campos del código HTML de una página.
EFECTIVIDAD DE LAS PALABRAS CLAVE	Grado de efectividad que se consigue al intentar posicionarse con una palabras clave. La efectividad medirá la popularidad de las palabras para los usuarios y el número de competidores con los que luchar para conseguir mejorar el posicionamiento.
RANKING TRÁFICO	Ranking de la página según el tráfico que recibe.
PAGERANK	Ranking de Google según la calidad de los enlaces de las páginas web.
ANÁLISIS ENLACES	Estudio de la morfología y diseño de los enlaces que apuntan a la página web propia.
FTP UPLOAD	Carga de las modificaciones de la página al servidor.
TEST DE VELOCIDAD	Test que mide la velocidad de carga de la página.
ALERTA DE PROBLEMAS	Alerta cuando se producen cambios en posicionamiento o caídas del servidor de la página.
CHEQUEO DIRECCIONES	Chequeo de las direcciones que pretenden intercambiar enlaces.
CONTENIDO DE LA PÁGINA	Estudio del contenido de la pagina en relación al total de la pagina.
FORMATO PALABRAS CLAVE	Estudio de los formatos de las palabras clave que pueden mejorar su valoración como: negrita, cursiva, mayúscula.
PAGERANK DC	Ranking de los enlaces de Google de cada uno de los datacenters que tienen distribuidos.
ANALIZADOR ETIQUETAS META	Estudio del diseño y aparición de las palabras clave en las meta_etiquetas.
SIMULADOR DE SPIDER	Visión de la pagina Web tal y como la ven los robots de indexación de los motores de búsqueda.
VISOR DE SNIPETTS	Vista de los Snipetts de la página.
VERSIONES ANTERIORES DE LA PÁGINA	Versiones guardadas anteriores de la página en Way Back Machine.
TAMAÑO DE LA PÁGINA	Número de bytes que ocupa la página web.

Tabla III-1: Funcionalidades de optimización web ofrecidas por las herramientas SEO

3.1.2.2 Correspondencias entre las herramientas SEO y las funcionalidades SEO

Una vez identificadas las funcionalidades de posicionamiento que ofrecen las herramientas SEO (ver Tabla III-1), se examina y se coteja nuevamente el estudio sobre aplicaciones SEO de la sección 2.2.3 y se establece la correspondencia entre las 21 herramientas principales y las funcionalidades SEO. Los resultados de este estudio se muestran en la Tabla III-2.

Por otra parte, las 18 aplicaciones SEO recogidas en la herramienta SEO Tools se desglosan aparte en la Tabla III-3. Estas herramientas tienen también identidad propia, es decir, son herramientas SEO en sí mismas que pueden funcionar de forma independiente.

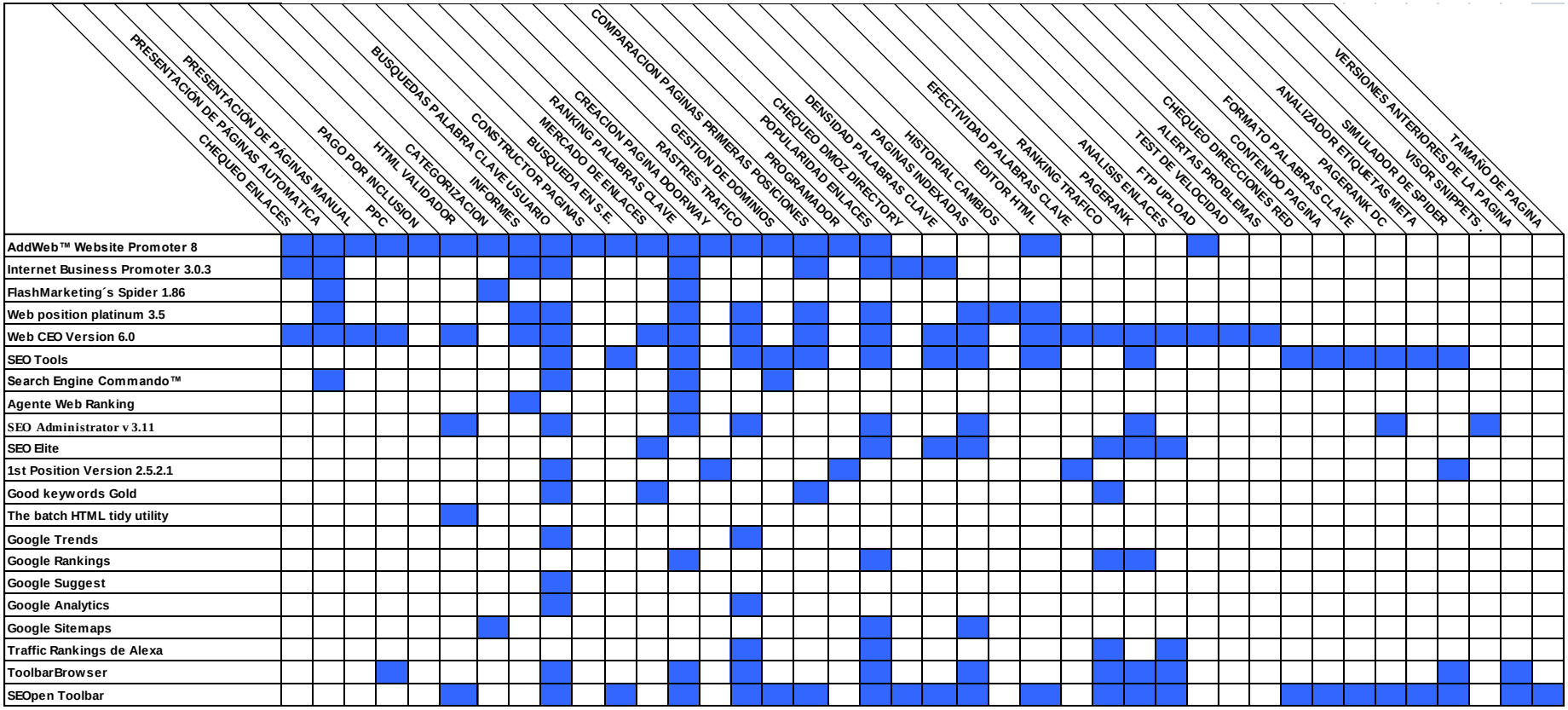


Tabla III-2: Correspondencias entre herramientas SEO y funcionalidades SEO

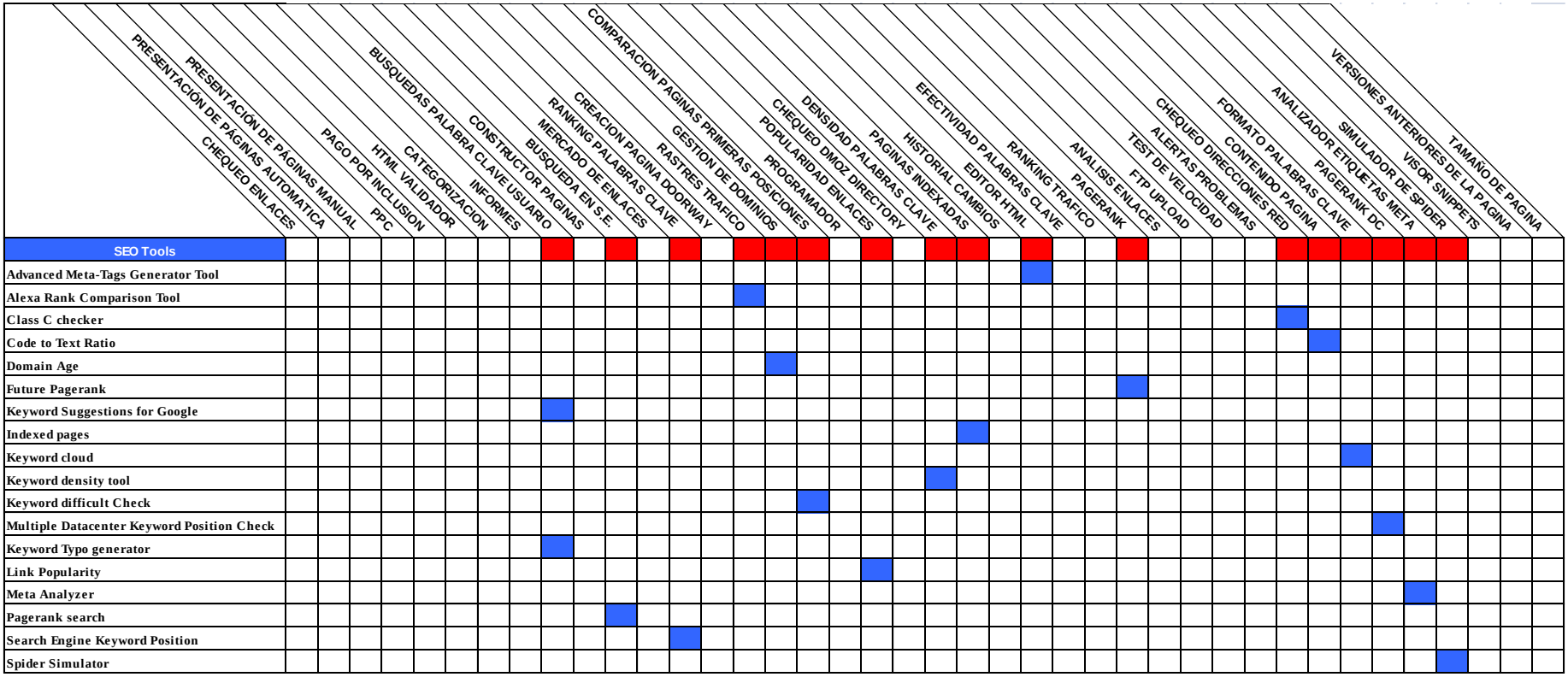


Tabla III-3: Correspondencias entre las aplicaciones integradas en la herramienta SEO Tools y las funcionalidades SEO

Para algunas funcionalidades se ha ampliado la información con las comparativas de las principales herramientas SEO que las ofrecen, según los campos de aplicación en los que actúan.

La comparativa destaca que la herramienta *Internet Business Promoter 3.0.3* recoge más campos que *Web Position Platinum*. Los campos que no tiene en consideración son el de *Comment* y el de *Body Text*.

Por el contrario la herramienta *Web Position Platinum 3.5* tiene en cuenta esos campos, pero no algunos relevantes como las etiquetas de los títulos <H1>, <H2> o la popularidad de los enlaces y algunos atributos relacionados con los enlaces.

En concreto, en la *Tabla III-4* se muestran los campos estudiados, por las herramientas *Internet Business Promoter 3.0.3* y *Web position platinum 3.5,* en la comparación con páginas que aparecen en las primeras posiciones de una consulta web.

CAMPO DE COMPARACION		***Internet Business Promoter 3.0.3***	***Webposition platinum 3.5***
Document title	Titulo del documento	■	■
Meta keywords	Etiqueta de palabras clave	■	■
Meta descripción	Etiqueta de la descripción	■	■
1st phrase in body	Primera frase del texto del cuerpo del documento	■	
All links text	Texto de todos los enlaces	■	■
Link popularity	Popularidad o importancia de los enlaces	■	
All links URL	Todas las URL de los enlaces	■	
Same site link URL	URL de los enlaces de pagina similares	■	
Outbounds links URL	URL de los enlaces externos	■	■
<H1> y <H2> headline texts	Texto de las cabeceras H1 y H2	■	
Same site link texts	Textos de los enlaces de sitios similares	■	
Outbound link texts	Textos de los enlaces externos	■	
HTML comments	Comentarios de HTML	■	■
Image Alt attributes	Características de las etiquetas ALT	■	■
Comment	Comentarios		■
Body text	Texto del cuerpo del documento		■

Tabla III-4: Campos de comparación en los primeros resultados de las consultas por las herramientas Internet Business Promoter 3.0.3 y Web Position Platinum 3.5

Y en la Tabla III-5 aparecen los campos de aplicación de los editores HTML de las herramientas *Web Position Platinum* 3.5 y *Web CEO Version* 6.0. En este caso la información se ha realizado con preguntas para saber si usan un determinado campo.

CAMPO DE COMPARACION		*Web position platinum 3.5*	*Web CEO Version 6.0*
Número máximo de repeticiones de la palabra clave en una fila			
¿Ha usado el mismo color para el texto y el fondo?			
¿Tiene áreas de entrada ocultas?			
¿Usa la etiqueta Meta refresh?			
¿Usa Frames?			
¿Usa controles?			
¿Usa Javascript?			
¿Usa VBScript?			
Setup	Configuración		
Meta Tags	Etiquetas meta		
Headings	Cabeceras		
Alt	Imágenes		
Links	Enlaces		
Content	Contenido		
Opcional Meta Tags	Etiquetas META opcionales		
Find & replace	Encuentra y reemplaza		

Tabla III-5: Campos de aplicación de los editores HTML de las herramientas: Web Position Platinum 3.5 y Web CEO 6.0

3.1.2.3 Comparación de herramientas SEO según el número de funcionalidades

Un análisis del número de funcionalidades de posicionamiento que presenta cada herramienta SEO desvela que los valores son muy dispares. Con la comparativa anterior (Tabla III-2) se obtienen la relación entre las aplicaciones SEO y la cantidad de funcionalidades que consideran. Destacan las herramientas *SEOpen Toolbar*, *Web CEO Version 6.0*, y *AddWeb™ Website Promoter 8* por el mayor número de funcionalidades ofrecidas y valores significativamente superiores. El resultado de los recuentos efectuados se muestra en la Figura III-3.

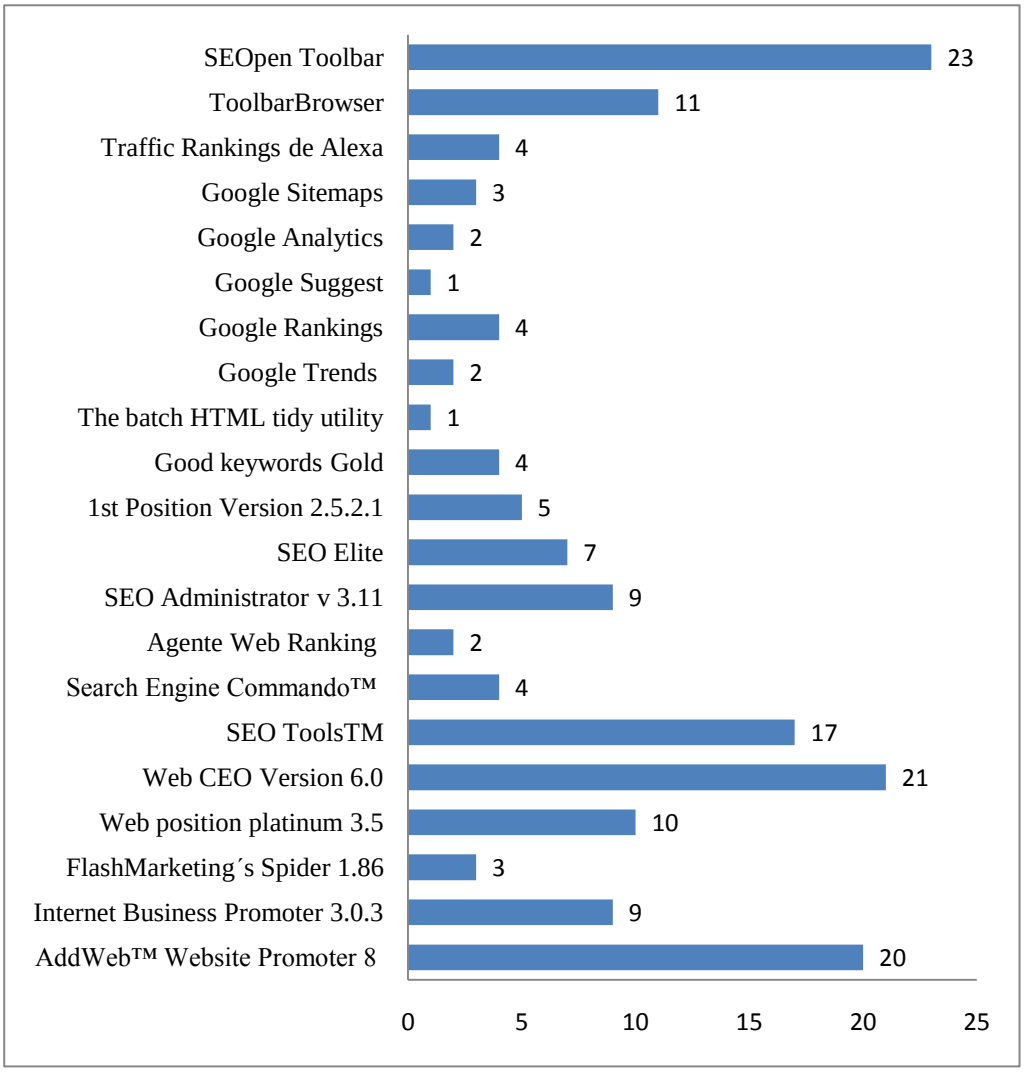


Figura III-3: Número de funcionalidades consideradas en cada herramienta SEO

3.1.2.4 Comparativa entre funcionalidades según el número de herramientas SEO que las aplican

Para cada funcionalidad considerada por las herramientas SEO (ver 2.1.7) se obtiene el número de herramientas que la utilizan a partir de la Tabla III-3. El resultado de la comparativa se presenta en la Figura III-4. La funcionalidad más extendida es "búsqueda de palabras clave" presente en un 66.6% de las herramientas SEO principales, seguida de "Ranking de palabras clave", y "popularidad de enlaces", ambas con un 57,1%.

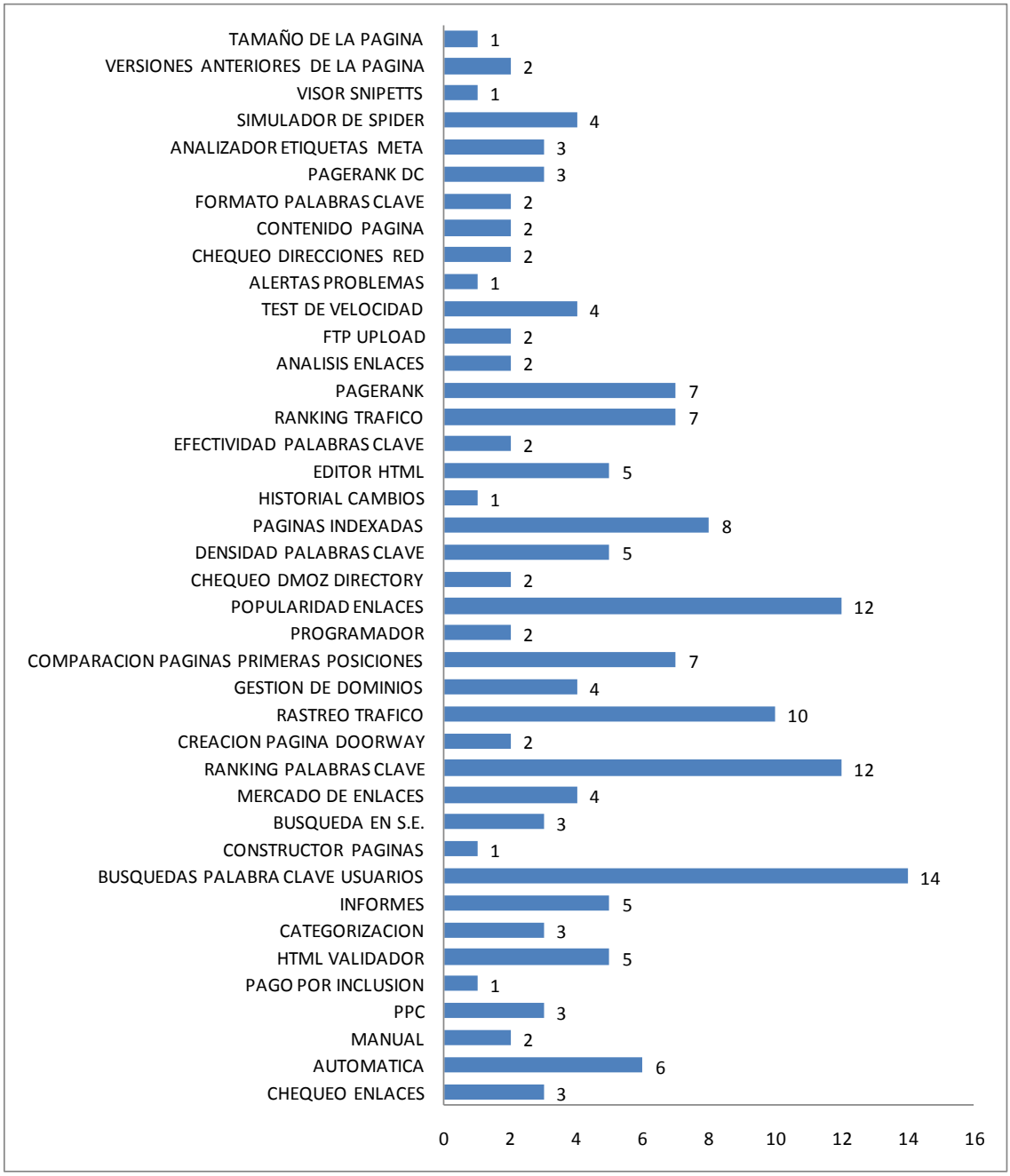


Figura III-4: Número de herramientas que estudian cada uno de los factores de posicionamiento

3.1.3 Discusión y conclusiones

Con estos estudios se ha identificado que los factores SEO más relevante utilizados por las herramientas SEO son la búsqueda de palabras clave de los usuarios, el ranking de palabras clave, la popularidad de los enlaces, el rastreo del tráfico, el número de páginas indexadas, la comparación de páginas en primeras posiciones, el ranking del tráfico y el pagerank. Los factores de posicionamiento más influyentes están relacionados con el contenido temático en primer lugar y después con la popularidad.

Destaca que el PageRank no es la que más peso tiene en el posicionamiento a pesar de que popularmente se considera el factor más determinante.

Por último, la identificación de los factores utilizados en la optimización de recursos web, su importancia atendiendo a las herramientas que los analizan y la cantidad de factores que consideran las herramientas más completas, constituyen la base en la que se apoya el siguiente bloque de trabajo.

3.2 Estimación de la relevancia documental asignada por los buscadores

El objetivo que se persigue es la construcción de un modelo que pueda decidir sobre la relevancia documental de una página web ante una determinada situación de búsqueda. Es decir, una vez establecidos el motor de búsqueda y las palabras clave de la consulta determinar cuál será la posición que alcanzará una página en el ranking de posicionamiento.

La pretensión perseguida es que el modelo o modelos construidos se comporten de forma análoga al buscador. Esto es que la relevancia que otorguen los modelos a las páginas web resultantes de una consulta concreta se corresponda con el posicionamiento realizado por el motor de búsqueda.

Para desarrollar esta fase se establece una metodología para la estimación de la relevancia documental, y se diseñará una experimentación de la que se expondrán y discutirán los resultados obtenidos.

3.2.1 Metodología para la estimación de relevancia documental

La generación de un modelo predictivo del posicionamiento web se aborda en tres etapas.

1. En la **primera etapa**, se generan clasificadores binarios para determinar cuál es la página más relevante para una determinada consulta en un buscador. Se aplican técnicas de aprendizaje automático sobre pares de resultados correspondientes a la consulta.
2. En la **segunda etapa**, se predice la posición de una página web (ya sea nueva o modificada) entre los resultados de la consulta analizada. Para ello, se aplica alguno de los clasificadores binarios construidos en la etapa anterior, comparando la página cuya posición se pretende estimar con cada una de los resultados de la consulta. A partir de los resultados de las comparaciones se infiere de forma automática la

posición que ocuparía el documento si estuviese incluido entre los resultados devueltos por el buscador.

3. En la **tercera etapa** se realiza la evaluación de la estimación de relevancia documental.

En definitiva, el conjunto de las etapas proporciona un emulador de los motores de búsqueda que se comporta de modo similar al ordenar los resultados por su relevancia. Se describen a continuación los pasos seguidos en la metodología y se presentan organizados en el diagrama de la Figura III-5

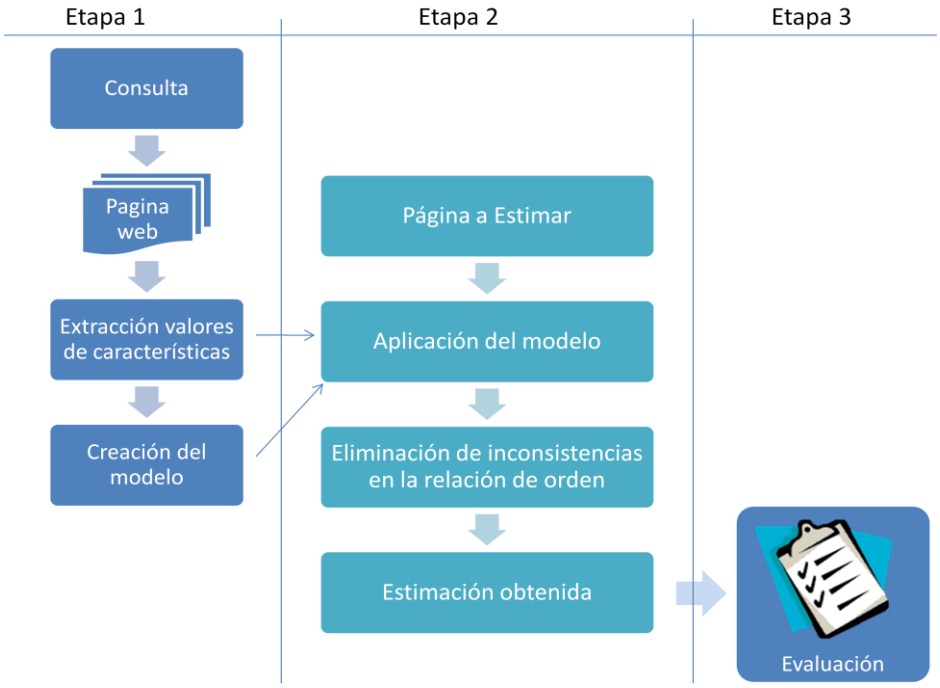


Figura III-5: Diagrama de flujo del proceso metodológico

3.2.1.1 Etapa 1: Generación de clasificadores binarios

La generación de clasificadores binarios sigue los pasos de selección de consultas, selección de páginas web, extracción de valores para las características, para finalmente poder crear el modelo.

- **Consultas**: En este paso, se elige un motor de búsqueda y se realiza la consulta o consultas sobre las que estudiar la relevancia web.
- **Páginas web**: Se tomarán como documentos web significativos las *n* primeras pág*i*nas web resultantes de cada consulta efectuada.
- **Extracción de los valores de las características**: Para cada una de las páginas web significativas se extraen de forma automática los valores de los factores de

posicionamiento que se hayan considerado interesantes y apropiados tras los estudios efectuados sobre variables SEO. La extracción automática se realiza con una aplicación software desarrollada para este fin.

- **Creación del modelo**: Se construye, con una herramienta de análisis de datos, un modelo por aprendizaje automático a partir de los valores de las características. Este modelo es capaz de decidir, al comparar dos páginas web, cual tiene mayor relevancia conforme a la consulta realizada. Resaltar que el modelo generado es binario, es decir, sólo actúa como discriminador frente a parejas de documentos web.

3.2.1.2 Etapa 2: Construcción de modelos de estimación a partir de clasificadores binarios

Para la construcción de modelos de estimación a partir de los clasificadores binarios se aplican repetidamente los modelos binarios y se eliminan las inconsistencias en la relación de orden.

- **Aplicación del modelo.**

  Este paso de la metodología responde a la necesidad de estimar la relevancia de un documento web en comparación con las de su competencia. Es decir, si se presenta una página web a un motor de búsqueda o se modifica una ya indizada ¿cuál sería la visibilidad asignada por el buscador ante una consulta determinada?

  Para responder a esta pregunta, en este paso, se compara la página en cuestión con cada una de las que constituirían su competencia mediante la aplicación reiterada de algún clasificador binario generado en la etapa anterior. El conjunto de resultados obtenido tras las comparaciones debe determinar la estimación de la posición.

- **Eliminación de inconsistencias en la relación de orden**:

  Este paso es producto de los posibles errores que pueden cometer los clasificadores binarios en las comparaciones. Si los modelos binarios generados fuesen infalibles, tal y como se aplican en el paso anterior, proporcionarían un conjunto de resultados de comparación compatibles con la relación de orden. Sin embargo, lo esperable es que algunos resultados se contradigan con otros incumpliéndose las propiedades antisimétrica y transitiva. En tal caso, se deben

corregir las decisiones erróneas o por lo menos estudiar formas de compensación de los fallos que permitan establecer estimaciones de posicionamiento.

- **Obtención de la estimación.**

  Por último, se aplica la estrategia de estimación seleccionada en el paso precedente. El resultado indica la posición que alcanzaría, el documento web en estudio, entre los de su competencia. Por tanto, el proceso concluye con una posición estimada de la visibilidad o relevancia documental.

3.2.1.3 Etapa 3: Evaluación de la estimación de relevancia documental

Para la evaluación del éxito de los estimadores se reservan resultados de las consultas, de manera que no intervengan en la construcción de los modelos, para posteriormente comparar las posiciones estimadas de estos documentos web con sus posiciones reales (las posiciones asignadas por el buscador). De esta forma se puede medir el grado de semejanza, entre el buscador web utilizado y el emulador construido, al otorgar relevancia a los resultados de una consulta. La validez de la metodología se puede contrastar a partir de los datos de las estimaciones.

3.2.2 Desarrollo de la estimación de relevancia documental

En este aparado se desarrollan los pasos expuestos en la metodología. La finalidad es obtener un modelo que permita estimar la posición que alcanzará un documento web entre un conjunto de documentos que representen su competencia ante una determinada consulta.

Se genera un modelo capaz de predecir la posición que una página adquiere en un determinado motor con anterioridad a que sea rastreada e indizada.

3.2.2.1 Etapa 1: Generación de clasificadores binarios

Para la generación de clasificadores binarios se seleccionan consultas y se extraen los valores de las características. Estos procesos están detallados y justificados a continuación.

I. Realización de consultas

La experimentación se ha realizado mediante consultas en motores de búsquedas. En concreto estos motores han sido Google, Yahoo Search y MSN. Las consultas correspondientes a cada experimento se han efectuado en lengua inglesa, dejando para

trabajos futuros las búsquedas en otros idiomas, aunque se presume independencia del idioma para las consultas.

Para cada consulta se han seleccionado los ciento cincuenta primeros resultados por considerar que es un volumen de datos suficientemente extenso para asegurar la calidad del aprendizaje automático, en el sentido de constituir un conjunto de datos significativo, y a su vez no tan amplio como para incluir páginas de mala calidad y por tanto poco competitivas.

Ejemplos de las consultas realizadas son:

- information analysis
- genetic algorithms
- green day
- new york airport

Como se puede observar son consultas expresadas por palabras clave tal y como habitualmente realizan los usuarios de motores de búsqueda. Las consultas han sido seleccionadas por su disparidad y distancia temática para evitar solapamiento de contenido. Se han seleccionado palabras compuestas, porque los usuarios no suelen preguntar por conceptos simples a menos que tengan un poder discriminatorio muy alto. Una palabra compuesta es más específica. Además algunos estudios de usuarios como el de López (2009) afirman que el número medio de términos de la consulta es superior a uno.

El número de consultas está limitado para asegurar que el algoritmo de ordenación de resultados no sufre cambios durante el proceso de captura de los datos.

II. Extracción de los valores de las características

Una vez obtenidas las páginas resultantes de una consulta se procede a extraer los valores de las características de posicionamiento que permiten medir la relevancia documental de una página web. Esto es, los valores de una serie de atributos fijados de antemano cuyo propósito es poder discernir la calidad de cada página en función de su visibilidad (posición que le ha otorgado el buscador entre las páginas obtenidas en esa consulta).

En la mayoría de las ocasiones se conocen cuáles son esas características y lo que se desconoce es cómo influyen en una ordenación por relevancia. En nuestro caso concreto los propios buscadores aportan información de las características o atributos que

intervienen en el posicionamiento. Además, foros y herramientas especializadas en optimización web (herramientas SEO) hacen públicos los parámetros que utilizan.

Por este motivo se ha realizado un estudio sobre las características empleadas en los buscadores Web y en las herramientas SEO más importantes (Ver apartado3.1).

A. Variables a analizar

Esta sección comprende todas las variables que se pueden extraer en la fase de experimentación del segundo bloque de cada sitio web devuelto como resultado de una consulta por un motor de búsqueda. La intención es evitar ambigüedades a la hora de entender que representa cada una de las variables y definir de forma concreta el cálculo de sus valores.

La terminología web se ha hecho tan popular que en muchas ocasiones se comprende mejor un término original, en inglés, que su correspondiente traducción al español. En este sentido, al nombrar los factores de posicionamiento se ha utilizado ambas lenguas en distintos apartados de esta tesis según se apreciase que uso podía ser más clarificador. Con el fin de asociar las parejas de denominadores se incluye la siguiente tabla:

FACTORES DE POSICIONAMIENTO ANALIZADOS	
Ingles	**Español**
Percentage of keywords in the URL	Porcentaje de palabras clave en la URL
Percentage of keywords in the title	Porcentaje de palabras clave en el título
Percentage of keywords in the body	Porcentaje de palabras clave en el BODY
Percentage of keywords in the links	Porcentaje de palabras clave en el texto de los enlaces salientes
Percentage of keywords in the h1 sections	Porcentaje de palabras clave en los títulos H1
Percentage of keywords in the h2 sections	Porcentaje de palabras clave en los títulos H2
Percentage of keywords in the h3 sections	Porcentaje de palabras clave en los títulos H3
Percentage of keywords in the p sections	Porcentaje de palabras clave en las secciones párrafo
Percentage of keywords in the alt sections	Porcentaje de palabras clave en las descripciones ALT de las imágenes

Percentage of keywords in the metakw section	Porcentaje de palabras clave en las " key words " de las etiquetas META
Percentage of keywords in the metadescription section	Porcentaje de palabras clave en la "descripción" de las etiquetas META
Percentage of linked pages containing the keywords	Porcentaje de las páginas vinculadas que contienen las palabras clave
Percentage of incoming links containing the keywords	Porcentaje de enlaces entrantes que contienen palabras clave
Percentage of broken links	Porcentaje de enlaces rotos
Number of frames	Número de frames
FACTORES DE POSICIONAMIENTO ANALIZADOS	
Español	**Español**
Number of images	Número de imágenes
Number of outgoing links	Número de enlaces salientes
Number of incoming links	Número de enlaces entrantes
Number of links to the same page	Número de enlaces internos
Number of links to the same domain	Número de enlaces salientes al mismo dominio
Number of links to other domains	Número de enlaces salientes a otros dominios
Position of the first keyword	Primera posición de las palabras clave en el cuerpo del texto
PageRank	PageRank
Is the webpage redirected?	Página redireccionada

Tabla III-6: Equivalencia idiomática entre los factores de posicionamiento analizados

- **Porcentaje de palabras clave en la URL**: Este parámetro mide el porcentaje de texto de la URL que contiene cualquiera de las palabras clave. El programa no tiene en cuenta la primera parte de la URL para calcular el porcentaje (es decir, el HTTP: / / o FTP: / /). En el caso de consultas realizadas mediante una frase exacta, todas sus palabras tienen que aparecer unidas en el mismo orden, sin espacios entre ellas.
- **Porcentaje de palabras clave en el título**: Este parámetro determina qué porcentaje del texto de la sección de título de la página web contiene alguna de las

palabras clave. En el caso de consultas realizadas mediante una frase exacta, todas sus palabras tienen que aparecer seguidas en el mismo orden. Tanto en este como en el resto de los parámetros relacionados con los porcentajes de palabras clave en una determinada sección del texto, sólo las palabras y los números se tienen en cuenta: otros símbolos y signos de puntuación son ignorados.

- **Porcentaje de palabras clave en el BODY**: Este parámetro mide el porcentaje de la sección del cuerpo de la página web que contiene las palabras clave. No sólo busca palabras clave en el texto de la página, sino también en todos los componentes que aparecen dentro de la misma, tales como párrafos, tablas, enlaces, etc. En las consultas mediante frases exactas sus términos tienen que aparecer juntos para ser contabilizados, y los signos de puntuación son ignorados.
- **Porcentaje de palabras clave en el texto de los enlaces salientes**: Esta variable calcula el porcentaje de palabras clave que aparecen en el texto de todos los vínculos salientes. Al igual que con las variables anteriores, las palabras clave que forman una frase exacta deben de aparecer unidas en el mismo orden para tenerse en cuenta, y los signos de puntuación son omitidos.
- **Porcentaje de palabras clave en los títulos H1**: En este parámetro se calcula el porcentaje de palabras clave que figuran en las secciones título H1 de la página web. El lenguaje HTML permite incluir secciones título dentro del cuerpo del documento, en diferentes niveles de jerarquía, siendo H1 el más elevado. Debido a esto, las palabras que aparecen en los títulos se supone que son de más importancia para el creador de la página web.
- **Porcentaje de palabras clave en los títulos H2**: Es similar al parámetro anterior, pero para las secciones título H2. H2 son los títulos de segundo nivel en la jerarquía de los títulos de texto, justo debajo en categoría de los títulos H1.
- **Porcentaje de palabras clave en los títulos H3**: Parámetro similar a los dos anteriores, pero para las secciones título H3. H3 son los títulos de tercer nivel en la jerarquía de los títulos de texto.
- **Porcentaje de palabras clave en las secciones párrafo**: En este parámetro se calcula el porcentaje de palabras clave que aparecen en las secciones de los párrafos del texto, que están marcados con la etiqueta `<p>`. Hay que tener en cuenta que también puede haber texto fuera de los párrafos, y algunas páginas

puede que ni siquiera tengan estas secciones. A priori, si una página web tiene <p> secciones, se supone que debe estar bien estructurada, y esto puede tener un efecto positivo sobre su posicionamiento.

- **Porcentaje de palabras clave en las descripciones ALT de las imágenes**: En esta variable se calcula el porcentaje de palabras clave de las descripciones de imágenes marcadas con la etiqueta ALT.
- **Porcentaje de palabras clave en las " keywords " de las etiquetas META**: Los documentos HTML pueden tener etiquetas meta que sirven para añadir más información sobre su contenido, sin que la información tenga que formar parte del contenido en sí. Parte de esa información son las *keywords* de las etiquetas META que contiene una lista de palabras clave relacionadas con la página web. Esta variable mide el porcentaje de palabras clave de la consulta contenidas en las *keywords* de las etiquetas META. Destacar que los webmasters pueden incorporar en las *keywords* de las etiquetas META tantos términos como deseen, por tanto, este parámetro es apropiado para controlar conductas contaminantes.
- **Porcentaje de palabras clave en la "description" de las etiquetas META**: Este parámetro mide el porcentaje de palabras clave en un aparatado similar al anterior. La diferencia es que la *description* marcada con la etiqueta META contiene una breve descripción del contenido de la página web, no sólo las palabras clave que están relacionadas con ella. El texto informativo puede ser tan largo como queramos.
- **Porcentaje de las páginas vinculadas que contienen las palabras clave**: Esta variable mide el porcentaje de las páginas vinculadas que contienen las palabras clave, siguiendo los enlaces salientes de las páginas web analizadas. Este atributo permite medir el grado de vinculación real de los recursos relacionados con el que se está analizando.
- **Porcentaje de enlaces entrantes que contienen palabras clave**: De la misma manera que es importante para un recurso web vincular a otros sitios web con un contenido similar, también es conveniente que esté relacionado por otros sitios web de temática afín. Esta variable representa el porcentaje de todos los enlaces entrantes que contienen al menos una ocurrencia de las palabras clave.

- **Porcentaje de enlaces rotos**: Es el porcentaje de enlaces no aptos para la navegación web.
- **Número de frames**: Esta variable simplemente cuenta el número de frames que componen toda la página.
- **Número de imágenes**: Esta variable cuenta el número de imágenes contenidas en la página web. Como una imagen se entiende una etiqueta HTML `<img>` que enlaza al archivo de imagen.
- **Número de enlaces salientes**: Este parámetro cuenta el número de enlaces salientes de la página web, es decir, el número de etiquetas HTML `<a>`. Esta variable no distingue entre el tipo de vínculos: pueden ser enlaces a otros sitios, al mismo sitio, etc.
- **Número de enlaces internos**: Este parámetro sólo tienen en cuenta los vínculos salientes que apuntan a una parte de la misma página web que está siendo analizada.
- **Número de enlaces salientes al mismo dominio**: Los enlaces salientes que se tienen en cuenta para esta variable son los que conducen al mismo dominio, es decir, aquellos cuya primera parte de la dirección es la misma que la de la página web analizada. Por ejemplo, si estamos analizando http://www.uc3m.es/home, cada enlace que comienza con http://www.uc3m.es será considerado como del mismo dominio.
- **Número de enlaces salientes a otros dominios**: Esta variable tiene en cuenta los vínculos salientes a dominios diferentes.
- **Número de enlaces entrantes**: Este parámetro cuenta el número de vínculos desde otras páginas web que enlazan con la página web analizada. Google proporciona una funcionalidad para saber qué sitios están apuntando a una dirección URL determinada. Debe tenerse en cuenta que al contabilizar estos enlaces no se comprueba coincidencias de palabras clave entre las páginas de origen y la analizada.
- **Primera posición de las palabras clave en el cuerpo del texto**: En esta variable se evalúa, respecto al número de palabras total, la precocidad en la aparición de palabras clave en la sección del cuerpo de la página web. Para el cálculo de este

valor se obtiene la media de las primeras posiciones alcanzadas por cada una de las palabras clave. Con el fin de adaptar la medida a las distintas longitudes de texto de los documentos se expresa el valor en forma de porcentaje. El porcentaje más alto (100%) se daría si una palabra clave aparece en la primera posición del texto. El valor más bajo (0%) se daría si la palabra clave aparece en la última posición. De esta manera, los valores de esta variable para las distintas páginas web pueden compararse fácilmente.

- **PageRank**: El PageRank es un valor dado por Google a cada página web que almacena en su base de datos, y es una medida de su calidad y su popularidad. El valor de PageRank actualmente oscila en un rango de 0 a 10, aunque no es habitual que una página alcance los valores extremos.
- **Página redireccionada**: Esta variable booleana comprueba si la página realiza una redirección automáticamente cuando accede a ella un usuario. A priori, a las páginas web, que tienen algún tipo de redirección se les presume falta de calidad.

B. La normalización de las variables

Con el fin de facilitar el análisis comparativo entre variables es necesario someter a algunas de ellas a un proceso de normalización al no estar acotadas superiormente, es decir no tienen una cota superior, pudiendo tomar valores descompensados. Por ejemplo, el número de enlaces que recibe una página web puede variar entre unidades y millares. Google resuelve la normalización del número de enlaces utilizando escalas logarítmicas (Moreno, 2005), también en este mismo artículo se normaliza el número de visitas a un recurso web con este tipo de escalas.

La idea es transformar cada valor x en $|\log_n x|$ (parte entera del logaritmo en base n del valor). En la experimentación, las funciones de $\log_2 x$ y $\log_3 x$ se han utilizado para normalizar los valores originales según se muestra en la siguiente tabla.

$\log_2 x$	
Intervalo del valor	**Valor normalizado**
[0, 2]	1
(2, 4]	2
(4, 8]	3

$\log_3 x$	
Intervalo del valor	**Valor normalizado**
[0, 3]	1
(3, 9]	2
(9, 27]	3

(8, 16]	4	(27, 81]	4
(16, 32]	5	(81, 243]	5
(32, 64]	6	(243, 729]	6
(64, ∞)	7	(729, ∞)	7

Tabla III-7: Escala logarítmica de normalización

Los valores pertenecientes a un intervalo adquieren el mismo valor normalizado. Este valor normalizado es el resultado de aplicar el logaritmo correspondiente al extremo superior del intervalo (salvo para el último intervalo).

Como puede verse, la tabla de normalización que utiliza $\log_3 x$ está dirigida a las variables que suelen tomar valores más altos. La siguiente tabla muestra qué variables se han normalizado con cada función logarítmica:

$\log_2 x$	nº de enlaces a la misma página
$\log_3 x$	nº de imágenes
	nº de enlaces salientes
	nº de enlaces entrantes
	nº de enlaces a otros dominios
	nº de enlaces al mismo dominio

Tabla III-8: Función logarítmica aplicada a cada variable

C. Consideraciones y decisiones en la extracción de los datos

La gran mayoría de páginas web son archivos HTML, estos archivos por su estructura y etiquetas permiten la captura automática de los datos asociados a los factores de posicionamiento listados en el apartado 2.1.7. Sin embargo, ante la presencia de otros tipos de archivos se ha tenido que decidir sobre la conveniencia de procesarlos. En otros casos, aún siendo archivos HTML, surgen complicaciones que también requieren un tratamiento especial. En este apartado se detallan las consideraciones tomadas en la extracción de las características.

- **La decodificación de los archivos PDF**

Se decidió dotar a los archivos PDF de una estructura similar a la de los archivo HTML, haciéndose así posible el análisis de su contenido de manera similar. Para ello, tras

identificar una dirección URL en formato PDF (debido a su extensión) se añaden las etiquetas HTML básicas.

```
<html>
<title> </ title>
<body> [contenido de texto sin formato del PDF] </body>
</html>
```

- **Páginas desechadas**
  - **Streaming:** Archivos como ASF (*Advanced Streaming Format*) o RM (*Real Media*) son comúnmente utilizados por las radios en línea, y por lo tanto ofrecen una secuencia de datos sin fin. El problema con estos archivos es que, como no tienen una duración específica, la aplicación destinada a la lectura de datos entraría en un bucle infinito.
  - **Archivos de tipo no procesable:** Algunos archivos tienen un formato no procesable, es decir, no son ni HTML ni PDF, y por lo tanto han de ser tratados como archivos de texto en bruto. El problema con estos archivos es que su contenido puede ser ilegible como texto sin formato, ya que pueden ser probablemente archivos multimedia.
  - **Archivos excesivamente grandes**: Otros archivos son muy grandes, en términos del tamaño que ocupan sus datos. Aunque fuese posible capturar el contenido de estos archivos, la probabilidad de que su contenido no sea HTML unido a que la mayoría de los servidores limitan la velocidad de descarga de los archivos hacen desaconsejable su tratamiento.
  - **Archivos HTML de tamaño desconocido**: Algunos servidores no informan sobre el tamaño de los archivos, por lo que se desconoce el tiempo a invertir en el proceso de extracción de los datos.

Los problemas expuestos tienen una forma común de resolverse: en el caso de detectar alguno de los problemas específicos se evita la lectura del contenido del archivo.

Para los dos primeros tipos de problemas, la manera más simple para detectar si se trata de contenido *streaming* es leer la cabecera HTML recibida desde el servidor, e ignorar cualquier contenido cuyo tipo MIME no sea "texto", como imágenes, aplicaciones, etc. (con la excepción de PDF, que son tratados de una manera especial).

En el caso de archivos demasiado grandes, se consulta su tamaño desde el servidor y se evita la lectura si es más grande que un límite previamente establecido. Este límite puede

ser configurado con el fin de ajustar el número de archivos posibles que pueden ser ignorados. De manera similar, también se puede ajustar el número máximo de líneas que se pueden leer de un archivo, para los casos en los que el tamaño de un archivo se desconoce a priori.

En total, el porcentaje de páginas no leídas por las decisiones adoptadas es relativamente bajo, como puede verse en la siguiente tabla, que muestra los porcentajes de páginas desechadas.

	Páginas desechadas		
Consulta	**Nº de páginas de tipo no procesable**	**Nº de páginas con excesivas líneas**	**Nº de páginas con excesivo tamaño**
Information analysis	9,94%	1,21%	0,79%
Green Day	10,39%	0,92%	0,55%
New York Airport	5,53%	0,71%	0,27%
Genetic algorithms	5,16%	0,68%	0,72%
Porcentajes totales	8,11%	0,88%	0,54%

Tabla III-9: Porcentaje de páginas descartadas

La mayoría de las páginas que se descartan son debido a tipo no procesable (es decir, transmisión de medios, archivos zip, imágenes, etc.), y tiene sentido no procesarlas, ya que su contenido no es analizable (tal y como se han definido las características de posicionamiento). No obstante, se juzga interesante un futuro estudio sobre el posicionamiento de los documentos web cuyos formatos no se ajustan al HTML ni al PDF.

- **Análisis de las páginas con varios frames**

  Aunque no es el diseño más común de una página web, a veces sucede que un sitio está compuesto por varios frames. Aparte de tener que leer varios documentos fuente HTML como frames, surge la necesidad de establecer una forma de medir los parámetros de este tipo de páginas web a partir de las medidas obtenidas en sus frames. Se establecen las consideraciones adoptadas:

- **Porcentaje de palabras clave en el título, URL y meta-descripción KW:** Estos porcentajes de palabras clave han de medirse en la página que contiene todos los frames.
- **PageRank**: Se considera el PageRank de la página principal ya que engloba el PageRanks de todos sus frames.
- **Número de imágenes y enlaces salientes**: Estas variables se calculan sumando los valores obtenidos para cada frame.
- **Número de enlaces entrantes**: Se ha considerado el número de enlaces entrantes relativos a la página principal, y por lo tanto debe ser calculado utilizando la URL de la página principal.
- **Posición de la aparición primera palabra clave**: Se calcula el promedio de las posiciones obtenidas a partir de la primera aparición de palabra clave en cada uno de los frames.

- **Páginas redireccionadas**

  Algunas páginas realizan una redirección a una URL diferente de la original. En esos casos, es necesario analizar el contenido de la página final, no de la original, ya que es lo que los motores de búsqueda realmente procesan.

  En los casos en que las etiquetas META "*refresh*" o "*url*" aparecerá en el encabezado del documento HTML la dirección que figura en las etiquetas sustituye la URL original.

- **Detección de enlaces rotos**

  En algunos casos, una URL válida nos dirige a una página web inexistente. Esto puede ser debido a un colapso temporal del servidor o a que la página se ha eliminado. Otras veces sucede que la dirección URL contiene un error y apunta a una página web que no ha existido nunca, o que la dirección no es válida, ya que no pertenece a ningún protocolo válido (HTTP, HTTPS, FTP, etc.). También es posible que los enlaces apunten a sitios seguros (HTTPS), que en algunos casos son inaccesibles sin los permisos adecuados.

  Con el fin de detectar estos enlaces rotos es necesario advertir cuando se produce un error en el momento de la apertura de la URL. Estas detecciones se realizan por medio de controles de tiempo de espera.

- **Tiempo de espera para la conexión**: Establecer el tiempo máximo de espera para la conexión a la dirección URL.
- **Tiempo de espera de lectura**: Establece el tiempo máximo de espera para la lectura de la fuente de datos, una vez realizada la conexión.

Los tiempos de espera son configurables, siendo los valores por defecto de 15 segundos para ambos.

III. Creación del modelo

A partir de las características o variables de las ciento cincuenta primeras páginas se construye un clasificador por aprendizaje automático (ya se ha mencionado anteriormente porque se escogió el número ciento cincuenta, podría ser un número distinto de páginas siempre y cuando se respeten los motivos de su elección). El objetivo del clasificador es poder decidir entre dos páginas web cual es la más relevante ante cierta consulta. Se han utilizado algoritmos de inteligencia artificial supervisados, en concreto algoritmos de inducción de reglas por las ventajas expuestas en el Capítulo II:.

Los algoritmos de inducción de reglas con los que se ha experimentado son: C 4.5 y PART. También se ha experimentado combinando clasificadores homogéneos para mejorar la precisión mediante las técnicas *Bagging* y *Boosting*, tomando como base los algoritmos anteriores. Otros tipos de algoritmos como redes de neuronas, máquinas vectoriales, algoritmos genéticos, etc. también podrían ser apropiados para la construcción del modelo, sin embargo se han seleccionado los algoritmos de inducción de reglas porque presentaban más ventajas iniciales.

Las técnicas de inducción de reglas reciben la información como un conjunto de casos (ejemplos de aprendizaje). Estos ejemplos se representan por un conjunto de atributos común que incluye el atributo de clase. Los valores de estos atributos distinguen unos casos de otros. A partir de los datos de entrada generan un árbol de decisión o un conjunto de reglas que proporcionará la clasificación de los nuevos ejemplos.

En nuestro caso, cada instancia o ejemplo de aprendizaje, independientemente del algoritmo utilizado, se construye a partir de cada pareja de páginas que se pueden formar con 100 de entre las 150 primeras páginas de una consulta. Por tanto el número de instancias de aprendizaje máximo será $V_{100,2}$ (variaciones de cien elementos tomados de dos en dos). Los atributos de cada instancia serán las características o variables de la

primera página más los de la segunda página más el atributo de clase. Este último atributo es el que contienen la información de cuál de las dos páginas es más relevante y se obtiene a partir de las posiciones que ocupan en los resultados de la consulta.

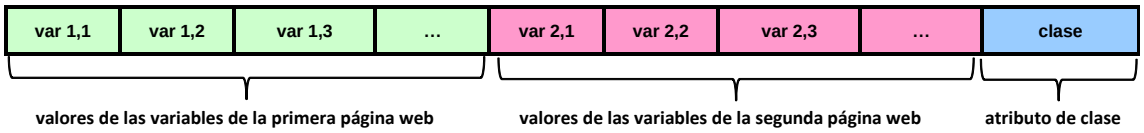


Figura III-6: Estructura de una instancia

La validación del modelo se efectúa con el método de validación cruzada (sobre diez subdivisiones estratificadas). Esta validación estima el comportamiento del modelo ante datos (páginas web) que no han formado parte del conjunto de aprendizaje, es decir, nos indica el porcentaje de acierto que tendrá el clasificador al decidir entre dos páginas cual es la más relevante para una consulta concreta.

3.2.2.2 Etapa 2: Construcción de modelos de estimación a partir de clasificadores binarios

Para la construcción de un modelo de estimación a partir de clasificadores binarios se debe aplicar un modelo binario y eliminar o compensar las inconsistencias en la relación de orden de las páginas. Estos procesos se presentan como una continuación de la etapa anterior.

IV. Aplicación del modelo

Si se desea saber la posición que un motor de búsqueda otorgará a una página web ante una determinada consulta se aplicará el modelo repetidamente para obtener la relevancia de dicha página frente a cada una de su competencia (las que aparecen como resultado de esa consulta).

Destacar que el modelo sólo necesita los valores de los atributos de esa página, por tanto, se puede estimar la relevancia de ese documento sin necesidad de que esté indexado. Asimismo se podrá saber el impacto que un documento web sufrirá en su posicionamiento ante hipotéticas modificaciones.

V. Eliminación de inconsistencias en la relación de orden

Como los modelos obtenidos por aprendizaje automático difícilmente son precisos en un 100%, podrían surgir inconsistencias respecto de la definición de relación de orden, ya que podrían incumplirse las propiedades antisimétrica y transitiva. Es decir, podría ser que, por ejemplo, se estimase que la página 1 es mejor que la 2, la 2 mejor que la 3 y la 3

mejor que la 1. Estos defectos o errores de precisión del modelo se subsanan mediante estadística, corrigiendo aquellas predicciones cuyo resultado sea más improbable, o de menor confianza, bien por estar en discordancia con respecto al resto de las predicciones, o bien corrigiendo la predicción más incoherente respecto al resto de las predicciones. Para esto se utilizan métodos estadísticos a partir de matrices de confusión, probabilidades de las decisiones, etc.

Por ejemplo, se puede optar por corregir el mínimo número de predicciones que permita establecer la relación de orden. También se puede reforzar esta idea teniendo en cuenta las probabilidades de decisión que cada regla del modelo tiene asociada.

Otra posibilidad, que se ha tenido en cuenta, consiste en admitir estas inconsistencias y aplicar políticas de compensación de errores que permitan corregir las hipotéticas desviaciones de las estimaciones. Por ejemplo, se puede aplicar el porcentaje de error que se comete al clasificar la página web a estimar, como más relevante que otra y viceversa. De esta forma si se han clasificado *K* páginas como más relevantes que el documento objetivo de la estimación, en lugar de asignarle directamente la posición *K+1*, se tendrán en cuenta los porcentajes de error mencionados.

3.2.2.3 Etapa 3: Evaluación de estimación de la relevancia documental

VI. Evaluación de la estimación de relevancia documental

La validación de la metodología se ha realizado reservando páginas de los resultado de las consultas, es decir, las páginas que se han reservado no han intervenido ni en la creación del modelo ni en las fases de eliminación de inconsistencias. Estas páginas web permiten medir la precisión a la hora de estimar el posicionamiento de una página. Comparando la posición estimada y la posición real (posición en la lista de resultados) de cada una de las páginas reservadas se obtiene la media, la desviación del error y el error máximo que se comete (error en nº de posiciones).

Los valores de estas medidas permiten precisar el éxito del estimador al predecir la relevancia documental.

3.2.3 Experimentación y resultados

Los experimentos desarrollados tienen como finalidad el estudio de la relevancia documental en la Web. El objetivo es poder decidir que documentos son los más

relevantes en un tema concreto, así como poder explicar los fundamentos de tales decisiones.

Para la experimentación y posteriores aplicaciones de esta investigación se ha desarrollado una aplicación el lenguaje Java (Knudsen y Niemeyer, 2005) cuyas funcionalidades permiten automatizar las tareas descritas en la metodología y desarrollo de presente bloque de investigación. Por tanto, permite:

- Capturar los ***n*** primeros resultados de una consulta o conjunto de consultas en los buscadores Google, Yahoo Search y MSN.
- Extraer los valores de los factores de posicionamiento de cada resultado.
- Construir instancias de cada posible par de parejas de resultados para generar modelos binarios.
- Generar y evaluar modelos binarios, a partir de la herramienta Weka, con las instancias construidas aplicando los algoritmos C4.5, PART y las técnicas de conjuntos de clasificadores *Boosting* y *Bagging* tomando como base estos algoritmos.
- Estimar la posición de un documento web entre un conjunto de resultados de una consulta.
- Evaluar el estimador de posición a partir de los resultados de las estimaciones.

3.2.3.1 Experimento 1: Construcción de emuladores para el motor de búsqueda Google

Como se ha explicado, varios algoritmos de inducción de reglas se han integrado en la aplicación desarrollada para realizar los experimentos, a fin de generar modelos de clasificación que sirvan para predecir las posiciones de las páginas web.

Los algoritmos seleccionados, se han elegido frente a otras opciones, porque son significativamente mejores en algunos aspectos como la robustez ante el ruido (la omisión o de falta de algún dato), identificación de atributos irrelevantes, extracción de reglas fáciles de entender y de gran expresividad, la posibilidad de modificar las reglas por los expertos, etc.

El objetivo de este experimento es averiguar cómo se comportan los algoritmos seleccionados al estimar la relevancia de un documento web y comprobar cuál es el más adecuado.

El experimento se realiza sobre los resultados de búsqueda que se han obtenido en Google para las siguientes cuatro consultas:

- Information analysis
- Green Day
- New York Airport
- Genetic algorithms

De cada una de estas consultas se recuperan los 150 primeros resultados de los cuales 100 se emplean en la construcción del modelo y el tercio restante se reserva para validar los estimadores generados. Los 50 resultados reservados por consulta se seleccionan al azar de entre las 150 páginas correspondientes conservando la información de la posición que ocupan entre los 100 resultados no seleccionados.

Tras extraer los valores de las características SEO de todos los resultados se construyen las instancias de aprendizaje. Para cada uno de los cuatro conjuntos de 100 resultados (un conjunto por consulta) se crea una instancia a partir de cada pareja ordenada de páginas que se puede formar, es decir $V_{100,2}$ (variaciones sin repetición de 100 elementos tomados de dos en dos). De este modo, cada instancia estará formada por las características de dos páginas más un atributo de clase (ver Figura III-6). El atributo de clase indica cual de las dos páginas está mejor posicionada. En total, tomando las instancias obtenidas a partir de las cuatro consultas se dispone de 39.600 ejemplos de aprendizaje.

Se genera ahora, con todas las instancias creadas, un clasificador binario con cada uno de los seis algoritmos de inducción de reglas. Estos algoritmos se han utilizado con la configuración estándar que proporciona la herramienta Weka.

Estos modelos binarios predecirán entre dos páginas web cuál es la más relevante para una determinada consulta en función de sus características y su validez se ha comprobado por evaluación cruzada con diez divisiones estratificadas.

Con los modelos binarios, se compara cada página web reservada para la evaluación del modelo definitivo con cada uno de los otros 100 resultados (no reservados) que están asociados a la misma consulta. La posición estimada para una página web reservada, entre las 100 con las que se le compara, viene dada por *K+1*, siendo *K* el número de páginas clasificadas como más relevantes, que la página web objeto de estimación, por el clasificador binario aplicado.

A partir de las estimaciones de las 200 páginas reservadas (50 por consultas) y sus posiciones reales entre los resultados de cada consulta se establece la validez del modelo de estimación en la emulación del buscador Google. Las medidas consideradas con el fin de medir el éxito en las predicciones de la relevancia documental son:

- Media del error: $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
- Desviación estándar del error: $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}$
- Error máximo: $\boldsymbol{max\ (x_i),\ 1 \le i \le N}$

Siendo x_i el *i-ésimo* error, definido como la diferencia en valor absoluto entre la posición real y la estimada de la *i-ésima* página web valorada.

$$x_i = |posición\ real - posición\ estimada|$$

Una vez que se ha experimentado con todos los algoritmos ofrecidos por nuestra aplicación para generar un estimador de posición, se compara los resultados de cada uno de ellos, a fin de saber cuál es el más adecuado para resolver el problema. La tabla siguiente resume los resultados obtenidos con cada algoritmo con los cuatro conjuntos de variables de entrada para Google. Para cada algoritmo se muestra el porcentaje de acierto del modelo binario que se genera (obtenido por validación cruzada) y las tres medidas que evalúan el éxito del estimador definitivo de posicionamiento, indicadas arriba. Los resultados obtenidos en la Tabla III-10 sobre la comparativa de los algoritmos aplicados al motor de búsqueda Google, destaca el algoritmo *Boosting* PART con un 90,46% de acierto y un error medio y desviación típica del error de 0,57 y 1,37 respectivamente.

Algoritmo	% acierto clasificador	Error medio	Desviación estándar	Máximo error
C4.5	83,36	2,79	3,02	38
PART	87,20	1,62	2,54	39
Bagging C4.5	88,35	1,57	2,44	39
Bagging PART	92,01	0,93	2,20	39
Boosting C4.5	88,30	0,57	1,39	38
Boosting PART	90,46	0,57	1,37	37

Tabla III-10: Comparativa algoritmos Google

3.2.3.2 Experimento 2: Adaptación de la metodología a otros algoritmos de posicionamiento

Este experimento se ha diseñado con la finalidad de conseguir una base empírica con la que dilucidar si el enfoque metodológico propuesto es capaz de adaptarse a otros algoritmos de posicionamiento. Es decir, si es aplicable a otros motores de búsqueda o a modificaciones en sus algoritmos.

El experimento se ha realizado repitiendo el experimento precedente para los buscadores Yahoo Search y MSN. Los resultados para ambos motores de búsqueda se muestran en las tablas dispuestas a continuación:

Algoritmo	% acierto clasificador	Error medio	Desviación estándar	Máximo error
C4.5	80,14	3,78	4,10	48
PART	86,80	3,19	4,06	39
Bagging C4.5	84,99	2,37	3,08	45
Bagging PART	90,05	1,76	2,9	41
Boosting C4.5	86,78	1,04	1,77	31
Boosting PART	87,81	1,07	1,67	35

Tabla III-11: Comparativa algoritmos Yahoo Seach!

Algoritmo	% acierto clasificador	Error medio	Desviación estándar	Máximo error
C4.5	82,47	4,25	4,99	45
PART	83,83	3,60	4,97	37
Bagging C4.5	82,16	2,04	3,51	43
Bagging PART	92,16	1,71	2,73	37
Boosting C4.5	88,09	1,04	1,08	33
Boosting PART	90,70	1,95	1,99	35

Tabla III-12: Comparativa algoritmos MSN

3.2.3.3 Experimento 3: Análisis de la distribución del error medio por posición

El los experimentos previos se han obtenido resultados sobre la precisión de las estimaciones de relevancia documental que han realizado los modelos construidos. A pesar de disponer de estos datos no se puede concretar si el éxito de las estimaciones es dependiente de las posiciones estimadas. Es decir, se desconoce si los errores medios de estimación se distribuyen de forma uniforme entre las posiciones. Esta información es de gran interés pues podría ocurrir que los estimadores se comportarán adecuadamente en ciertos rangos de posiciones y acumularán todos sus defectos en otras zonas.

En este experimento se va a analizar la distribución del error medio de estimación entre las cien primeras posiciones de los resultados de las consultas. Siguiendo la misma filosofía de los experimentos anteriores y una vez elegido el buscador web se van a reservar 50 páginas al azar, pero esta vez de entre las cien primeras posiciones, para cada una de las cuatro consultas que hasta ahora se han utilizado en la experimentación. A diferencia de los otros experimentos las 50 posiciones seleccionadas serán las mismas en todas las consultas. Se construye un modelo de asignación de relevancia para cada uno de los seis algoritmos de aprendizaje. Cada uno de estos modelos se pone a prueba con las páginas reservadas y se calcula la media de los errores para cada una de las 50 posiciones.

Se repite este proceso intercambiando el conjunto de páginas reservadas por aquellas que fueron seleccionadas entre las cien primeras completándose así el experimento.

En la siguiente gráfica se muestra el error medio por posición que comete el estimador construido a partir del algoritmo *Boosting PART* al emular al buscador Google. Advertir que, aunque el cálculo de la media de error se ha realizado a partir de los valores absolutos de los errores, se muestran en la gráfica resultados negativos en los casos en que la cuantía de los errores negativos superaba a la de positivos. De esta forma en la misma gráfica se dispone simultáneamente para cada posición del error medio y de la tendencia del modelo a sobrestimar o subestimar.

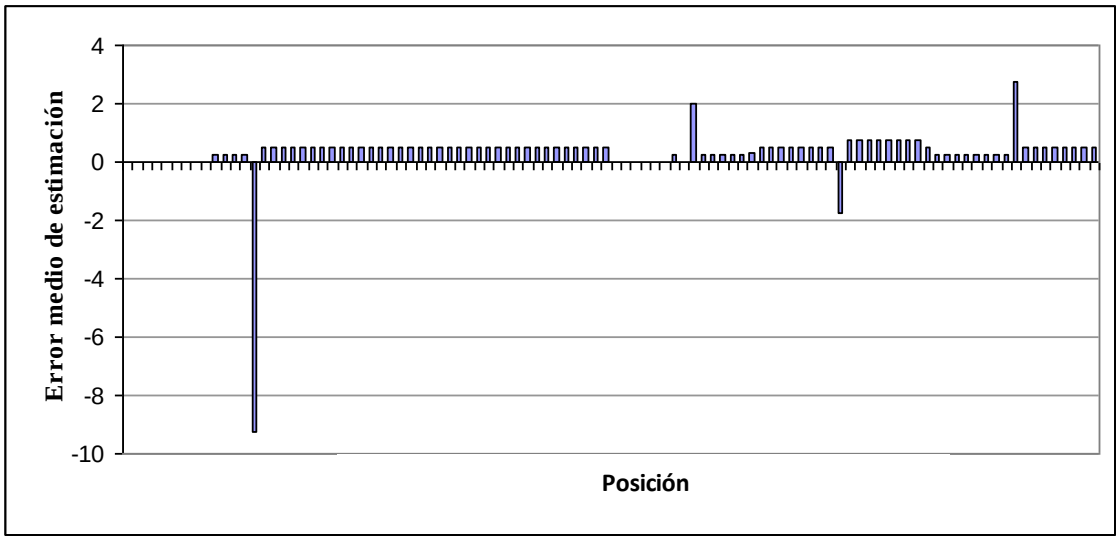


Figura III-7: Media del error por posición

3.2.3.4 Experimento 4: Comprobación de dependencia o independencia entre modelos y consultas

Con este experimento se pretende responder a la siguiente cuestión: ¿los modelos generados pueden estimar la relevancia documental atendiendo a un conjunto de palabras clave diferente del utilizado en las creaciones de esos modelos?

Expresado de otro modo, ¿se aprende el algoritmo genérico de un motor de búsqueda o la forma de posicionar documentos concernientes a consultas concretas?

Con motivo de responder a estas preguntas, se procede a efectuar una nueva consulta *"Computer engineering"* en los tres motores de búsqueda (Google, Yahoo Search y MSN). Para cada uno de estos buscadores se obtienen los 150 primeros resultados de la consulta y se extraen sus características de posicionamientos.

Un tercio de los resultados se aparta. Tras reservar estos resultados se estima su posición entre los 100 restantes mediante los modelos correspondientes (según el buscador utilizado) generados en los experimentos 1 o 2 (ninguno de estos modelos se genero a partir de la consulta *"Computer engineering"*).

Los resultados, para los tres buscadores, se muestran en las siguientes tablas indicando para cada modelo el porcentaje de acierto del clasificador binario al comparar dos páginas (de la nueva consulta) por su relevancia, y la media del error, la desviación típica y máximo error de las estimaciones de posición (midiendo el error en número de posiciones a partir de las ordenaciones proporcionadas por los buscadores web).

Algoritmo	% acierto clasificador	Error medio	Desviación estándar	Máximo error
C4.5	28,39	18,57	28,39	80
PART	28,9	20,68	28,9	87
Bagging C4.5	26,62	19,17	26,62	79
Bagging PART	30,02	22,01	30,02	84
Boosting C4.5	31,27	21,02	31,27	84
Boosting PART	28,36	19,27	28,36	82

Tabla III-13: Resultados para la consulta "Computer engineering" en Google

Algoritmo	% acierto clasificador	Error medio	Desviación estándar	Máximo error
C4.5	26,11	17,28	27,98	75
PART	29,45	21,31	28,2	84
Bagging C4.5	27,1	18,32	28,52	74
Bagging PART	31,52	21,44	32,22	82
Boosting C4.5	30,55	20,42	28,98	82
Boosting PART	29,54	21,28	29,66	84

Tabla III-14: Resultados para la consulta "Computer engineering" en Yahoo

Algoritmo	% acierto clasificador	Error medio	Desviación estándar	Máximo error
C4.5	26,32	17,78	28,41	73
PART	28,55	22,71	28,87	81
Bagging C4.5	28,17	20,20	28,98	79
Bagging PART	30,01	21,06	31,77	78
Boosting C4.5	31,22	20,44	29,02	88
Boosting PART	29,43	22,54	28,84	89

Tabla III-15: Resultados para la consulta "Computer engineering" en Msn

3.2.3.5 Experimento 5: Comparación de la eficiencia de los algoritmos

En el presente el experimento se obtienen medidas de los tiempos de ejecución de los algoritmos de aprendizaje empleados en este bloque de investigación. La comparativa de estas medidas permitirá establecer una valoración de sus rendimientos.

Con el fin de evaluar el comportamiento temporal de los algoritmos con respecto de la carga de entrada, se han dividido los resultados procedentes de las consultas sobre los buscadores en cuatro grupos. Por ejemplo, si nos centramos en las consultas que se realizaron sobre Google en el primer experimento, cada uno de los cuatro grupos contiene el 25 % de las instancias que representan a los resultados asociados a dichas consultas.

Para cada uno de los algoritmos se construyen 4 modelos, el primer modelo con un sólo grupo de instancias, el segundo con dos, y así sucesivamente.

Los resultados obtenidos, con las instancias correspondientes al Google, se muestran en la siguiente gráfica (para los otros dos buscadores no existen diferencias cualitativas):

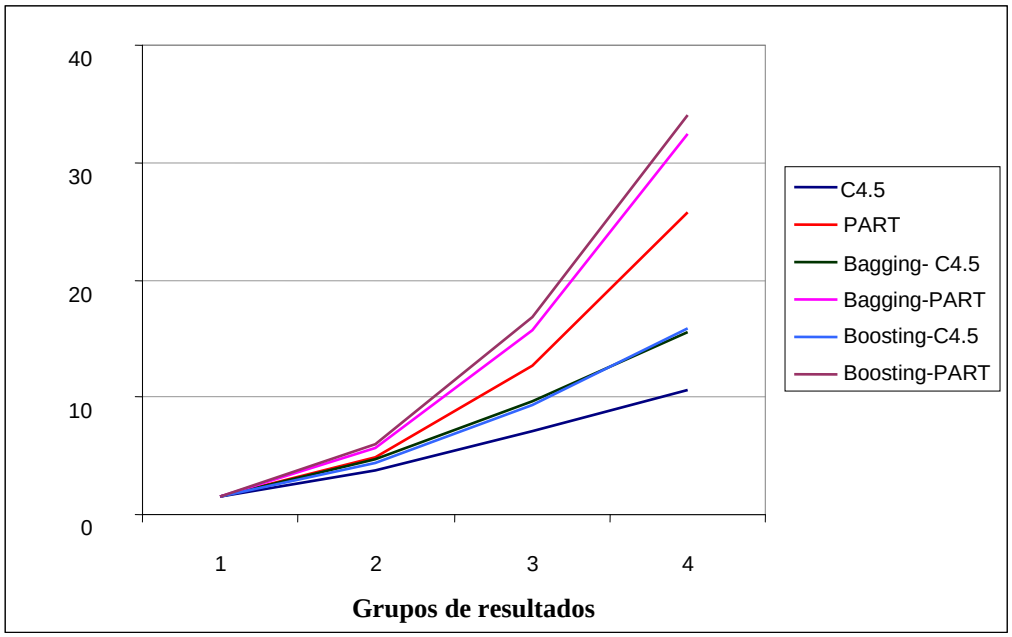


Figura III-8: Comparativa de tiempos de ejecución de los algoritmos

3.2.3.6 Experimento 6: Ajuste de la configuración de los algoritmos

En los experimentos anteriores los algoritmos se han aplicado con la configuración estándar de parámetros que proporciona la herramienta Weka. Este experimento está orientado a comprobar si es idónea esta configuración o por el contrario es conveniente afinar el ajuste de sus principales parámetros.

Existen varios parámetros configurables en los algoritmos, siendo los más importantes el número de iteraciones del algoritmo (número de clasificadores base generados en los algoritmos de conjuntos de clasificadores) y el factor de confianza (determina la poda a realizar en la generación de árboles de decisión).

La experimentación de ajuste del número de iteraciones se ha realizado sobre los cuatro algoritmos de conjuntos de clasificadores (los que utilizan las técnicas *Boosting* o *Bagging*). Estos experimentos se han realizado repitiendo los experimentos anteriores (1 y 2) para distintos valores del número de iteraciones (1,5,10,15,…,35) y fijando el factor de confianza de sus algoritmos base al 25% (valor estándar de Weka).

Para los experimentos relacionados con el ajuste del factor de confianza también se han replicado los experimentos 1 y 2 variando, en orden de decenas, los factores de confianza de los algoritmos C4.5 y PART, tanto en sus versiones propias como en sus participaciones como algoritmos base. En los algoritmos de conjuntos de clasificadores se ha seleccionado el mejor número de iteraciones obtenido al experimentar con este parámetro.

En las siguientes tablas se muestran, a modo de ejemplo, los resultados del ajuste de parámetros de uno de los experimentos preliminares (estos experimentos se realizaron con la intención de encontrar una configuración inicial con la que aplicar los algoritmos). Se marcan los mejores valores obtenidos para cada variable de medida del éxito en la predicción.

Iteraciones	% acierto clasificador	Error medio	Desviación estándar	Máximo error
1	82,40	3,83	4,36	40
5	86,07	0,95	2,85	37
10	87,17	0,98	2,84	38
15	87,70	1,00	2,85	38
20	87,74	0,97	2,86	40
25	87,75	1,00	2,84	37
30	87,75	1,00	2,84	37
35	87,75	1,00	2,84	37

Tabla III-16: Resultados para diferentes números de iteraciones

% factor confianza	% acierto clasificador	Error medio	Desviación estándar	Máximo error
10	86,38	1,04	2,88	42
20	86,09	1,03	3,04	44
30	85,91	0,99	2,86	39
40	87,74	0,98	2,91	43
50	85,81	0,98	2,86	39
60	84,96	1,01	2,97	42
70				
80				
90				
100				

Tabla III-17: Resultados variando el factor de confianza

3.2.3.7 Experimento 7: Construcción de modelo de estimación de la relevancia web (Boosting C4.5)

En los experimentos anteriores la técnica de conjuntos de clasificadores *Boosting* con algoritmo base *C4.5* ha destacado por su eficacia y eficiencia en la creación de modelos de estimación de la relevancia documental. En este experimento se pone a prueba esta alternativa de emulación de buscadores con una batería de consultas más amplia sobre el motor de búsqueda Google.

Las consultas se han tomado de forma semialeatoria a partir del autocompletable de la interfaz de Google. Por tanto, las consultas se corresponden con búsquedas frecuentes de los usuarios en el buscador Google. El proceso de selección de consultas se ha realizado introduciendo una por una las letras del alfabeto ingles en el cuadro de búsqueda de Google. Al introducir una letra (siempre sobre el cuadro de búsqueda vació) el buscador muestra alternativas de autocompletado de consultas de las que se han seleccionado cuatro al azar de entre las que tenían más de un término de búsqueda. De esta forma se han recopilado 104 consultas (26 letras * 4 alternativas).

En la siguiente Figura III-9 se muestra las alternativas de búsquedas que presenta la interfaz de Google al introducir la letra ‘a’.

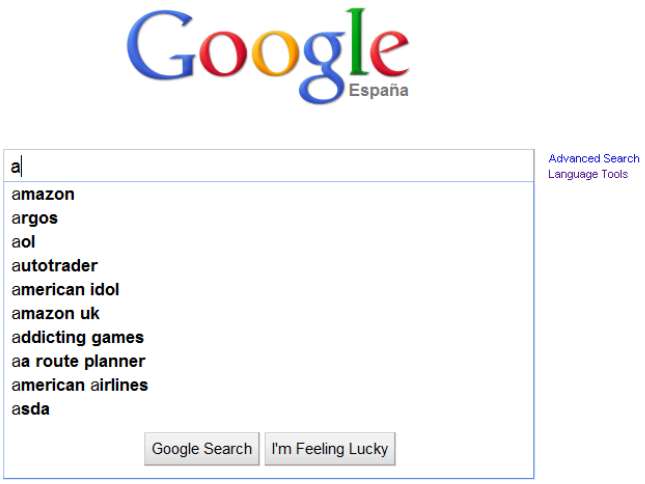


Figura III-9: Selección semialeatoria de consultas

De manera análoga al experimento 1 se realizan las consultas en Google y se capturan los 150 resultados de cada una de ellas, reservando la tercera parte (50 resultados por consulta) para la evaluación de los modelos.

En este experimento no se agrupan consultas para generar modelos, es decir, se construye un modelo por consulta. Por tanto, en la construcción de cada modelo intervienen los 100 resultados no reservados de la consulta correspondiente. Esto implica que en cada modelo binario intervienen 9.900 instancias ($V_{100,2}$), formadas tras extraer las características SEO de los resultados de la consulta.

Una vez obtenidos los modelos binarios con el algoritmo *Boosting C4.5* se compara con cada uno de ellos las 50 páginas reservadas para la validación con los 100 resultados utilizados en la creación del clasificador. Las diferencias entre las posiciones estimadas y las posiciones que realmente ocupan los resultados reservados son un indicador del error que se comete en los pronósticos de relevancia documental.

Los errores cometidos en la emulación del buscador, en este caso Google, se presentan mediante las medidas definidas en el primer experimento. Sin embargo, al repartirse los datos en grupos, uno por consulta, se dispone de la media y la desviación típica del error para cada uno de los 104 modelos de estimación generados, pero no de las medidas globales correspondientes. Se expresan a continuación las formulas que permiten hallar estas medidas globales a partir de las medidas particulares (Rodríguez-Miñón, 1982) de los 104 grupos.

La media, $\bar{x}$, de un conjunto de $\boldsymbol{n = n_1 + \ldots + n_r}$ observaciones se calcula como:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{r} n_i \bar{x}_i$$

siendo ***r*** el número de grupos, cada grupo *i* está formado por $\boldsymbol{n_i}$ observaciones de media $\bar{x}_i$

Como todos los modelos se han evaluado con el mismo número de datos se tiene que $n_i = \frac{n}{r}$ por tanto:

$$\bar{x} = \frac{1}{r}\sum_{i=1}^{r} \bar{x}_i$$

La varianza, $\boldsymbol{S^2}$, de un conjunto de $\boldsymbol{n = n_1 + \ldots + n_r}$ observaciones se calcula como:

$$S^2 = \frac{1}{n}\sum_{i=1}^{r} n_i S_i^2 + \frac{1}{n}\sum_{i=1}^{r} n_i (\bar{x}_i - \bar{x})^2$$

siendo ***r*** el número de grupos, cada grupo *i* está formado por $\boldsymbol{n_i}$ observaciones de media $\bar{x}_i$ y de varianza $\boldsymbol{S_i^2}$. Aplicando de nuevo la igualdad $n_i = \frac{n}{r}$ se tiene:

$$S^2 = \frac{1}{r}\sum_{i=1}^{r} S_i^2 + \frac{1}{r}\sum_{i=1}^{r} (\bar{x}_i - \bar{x})^2$$

Finalmente la desviación típica se expresa como:

$$S = \sqrt{\frac{1}{r}\sum_{i=1}^{r} S_i^2 + \frac{1}{r}\sum_{i=1}^{r} (\bar{x}_i - \bar{x})^2}$$

Los resultados globales obtenidos después de aplicar las formulas a los resultados de media y desviación típica de los 104 modelos se presentan en la siguiente tabla junto al máximo error cometido.

Motor de búsqueda	Algoritmo	Error medio	Desviación estándar	Máximo error
Google	**Boosting C4.5**	0,84	1,53	43

Tabla III-18: Resultados conjuntos de modelos de estimación (Boosting C4.5)

3.2.4 Discusión y conclusiones

Se ha evaluado la asignación de relevancia de los motores de búsqueda web a través de la posición que han ocupado los resultados de una consulta para la construcción automática de estimadores de relevancia documental. Se han realizado modelos de aprendizaje del comportamiento de los buscadores web tratando de emular el algoritmo de posicionamiento de los motores de búsqueda.

En los resultados obtenidos destacan los modelos de estimación construidos a partir de algoritmos de inducción de reglas de los conjuntos de clasificadores *Boosting* y *Bagging*. Con el algoritmo de *Bagging* PART se ha obtenido un porcentaje de acierto del 90% al decidir entre dos páginas cual es la más relevante en función de los factores de posicionamiento de los buscadores. No obstante a la vista de los resultados obtenidos con los algoritmos de Boosting se establecen como los modelos definitivos. Estas técnicas adoptan un mecanismo de compensación de errores en el cálculo de las posiciones estimadas por lo que obtiene mejores resultados.

3.3 Determinación automática de la influencia de cada factor en los algoritmos de ordenación de los motores de búsqueda

En este último bloque de investigación se identifican los factores que más influyen al posicionamiento en los buscadores Google, Yahoo Search y MSN. Para este propósito se aplican métodos de selección de atributos sobre el conjunto de factores de posicionamiento utilizado en el bloque de investigación anterior (3.2).

3.3.1 Metodología para la determinación automática del grado de influencia de los factores de posicionamiento

Los pasos seguidos en la metodología aplicada en la identificación de los factores más determinantes en la relevancia web asociada a un buscador son:

1. **Elección del algoritmo de selección de atributos.** En la búsqueda de los factores más decisivos para el posicionamiento web han de descartarse los algoritmos de selección de atributos basados en evaluaciones de subconjuntos de atributos. Estos algoritmos prescinden de los atributos redundantes, es decir, seleccionan grupos de atributos con poca interrelación entre ellos. De aplicarse se corre el riesgo de no obtener una medida objetiva de la importancia individual de cada uno de los

factores de posicionamiento. Por consiguiente, se debe aplicar algoritmos de selección que evalúen individualmente los atributos y los ordenen por su calidad.

2. **Comprobación de la validez del método de selección.** La selección de atributos además de incluir una búsqueda, con la estimación de la utilidad de cada atributo, se completa con una evaluación respecto a un esquema de aprendizaje específico (Hall y Holmes, 2002). En concreto, se generarán modelos binarios de forma análoga, excepto porque sólo estarán involucrados los atributos seleccionados, a los creados en el bloque de investigación previo. La comparación de los modelos obtenidos a partir de la selección de atributos con aquellos que los contemplan todos aportara indicios de la validez de la selección.
3. **Evaluación del impacto por la reducción de atributos en la estimación de la relevancia Web.** Los clasificadores binarios construidos en el punto anterior, con los atributos seleccionados, se emplean en la construcción de modelos de estimación de relevancia siguiendo la metodología propuesta en 3.2.1. Comparando la validez de sus pronósticos con las predicciones de los modelos alimentados con el conjunto total de atributos se deduce el efecto causado por la reducción.

3.3.2 Desarrollo

Los puntos tratados en la metodología con el propósito de esclarecer el peso de los factores de posicionamiento en los algoritmos de ordenación de los buscadores se desarrollan en este apartado.

3.3.2.1 Elección del algoritmo de selección de atributos

Como se ha explicado en el punto correspondiente de la metodología se debe utilizar un algoritmo de selección que evalué los atributos de forma independiente.

En concreto se ha elegido el algoritmo ChiSquaredAttributeEval que evalúa los atributos individualmente y que aplicado junto al método Ranker nos proporciona, en este estudio, una ordenación de los factores de posicionamiento según su importancia.

Los datos se presentan conforme a la experimentación del bloque de investigación precedente (ver Figura III-6), es decir, ante dos resultados de una misma consulta (fijado el buscador) su representación en forma de instancia viene dada por los valores de los atributos de posicionamiento de una de las páginas, a continuación, los valores de los

atributos de la otra página y finalmente el atributo de clase indicando la página que tiene mejor posición.

Por tanto, el algoritmo ChiSquaredAttributeEval determina en este caso la correlación entre cada factor de posicionamiento, en la comparativa de dos páginas, y la relevancia entre ellas (según el buscador) a partir del valor estadístico Chi-cuadrado (Freund et al., 2000).

3.3.2.2 Comprobación de la validez del método de selección

Llegados a este punto se dispone de la lista de factores de posicionamiento ordenados según su utilidad en la clasificación. Según se ha explicado en el punto anterior, las instancias de entrada del algoritmo de selección son las utilizadas en el segundo bloque de investigación en la comparación de parejas de páginas web. Por tanto, la validez de las selecciones mediante esquemas de aprendizaje (Hall y Holmes, 2002) se debe realizar por comparación de clasificadores binarios. Es decir, comparando los clasificadores binarios obtenidos en el bloque anterior con todos los atributos, con los que se construyan con los atributos seleccionados.

Los pasos a seguir para la comprobación de la validez del método de selección son:

- **Distinguir en el ranking de atributos los factores más importantes.** Recordar que el método ChiSquaredAttributeEval no proporciona un subconjunto de atributos, siendo necesario establecer un punto de corte en la lista ordenada.
- **Construir con los atributos seleccionados clasificadores binarios.** Estos clasificadores se generan de manera semejante a los construidos en el bloque de estimación de la relevancia, coincidiendo algoritmos de aprendizaje e instancias, salvo por la reducción de los atributos.
- **Comparación de la eficacia de los clasificadores.** Tras evaluar los clasificadores binarios creados con los atributos seleccionados se comparan los resultados examinando si la reducción del número de atributos afecta significativamente a la eficacia de los clasificadores.

3.3.2.3 Evaluación del impacto por la reducción de atributos en la estimación de la relevancia web

En la metodología del bloque 2 los clasificadores binarios son el pilar principal de los modelos de estimación de la relevancia documental web. Estos modelos de estimación se

evalúan atendiendo a mediadas basadas en los errores cometidos. Repitiendo este proceso con los clasificadores binarios generados con los factores seleccionados se obtiene por comparación otra medida de la calidad de la selección.

La importancia de cada factor de posicionamiento varía dependiendo del motor de búsqueda, sin embargo, debe ser independiente de las consultas que se realicen. Por ello, en los experimentos de aprendizaje orientados a evaluar los pronósticos de relevancia se incorporan grupos de instancias de distintas consultas.

3.3.3 Experimentación y resultados

Los experimentos realizados en esta etapa están orientados a la identificación de los factores de posicionamiento más significativos en los buscadores Google, Yahoo Search y MSN. Además, algunos experimentos se han enfocado para poder validar la metodología aplicada en las selecciones de atributos. En estos experimentos se construyen modelos de estimación del posicionamiento, para el buscador Google, a partir de los factores seleccionados. La comparación de los resultados con los obtenidos en el bloque de investigación precedente determinará el impacto de la selección.

Las funcionalidades de la aplicación desarrollada para la experimentación se han ampliado de forma que, además de indicar los factores más significativos, se puedan realizar los experimentos del bloque segundo de investigación tras seleccionar un subconjunto de atributos.

3.3.3.1 Experimento 1: Factores más influyentes en las ordenaciones de resultados de Google

En este experimento se aplica el método de evaluación de atributos ChiSquaredAttributeEval junto al método de búsqueda Ranquer a la colección de factores de posicionamiento listados en la Tabla III-6. Ambos algoritmos también están incluidos en el repertorio de la herramienta Weka.

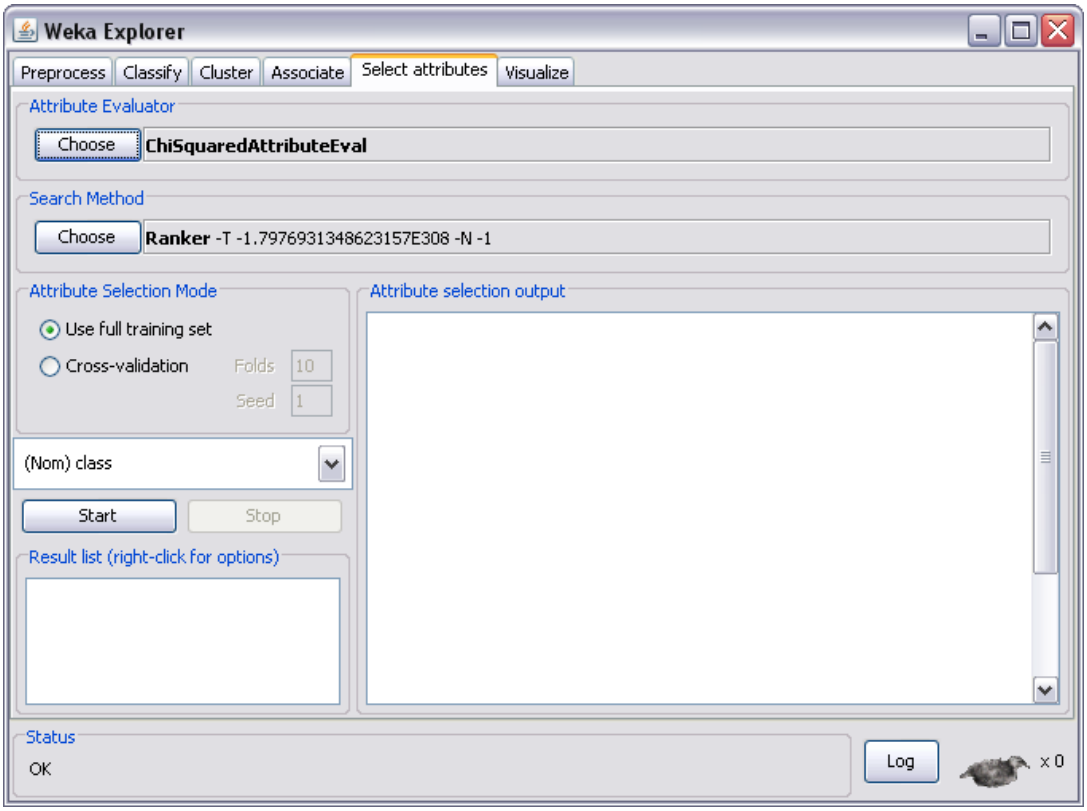


Figura III-10: Selección de atributos en herramienta Weka

Las instancias de entrada son las mismas con las que se construyeron los modelos binarios para el buscador Google en el primer experimento del anterior bloque de investigación (apartado 3.2.3). Por tanto, al intervenir una pareja de páginas en cada instancia cada factor de posicionamiento aparece dos veces (ver Figura III-6). Esto implica que en la lista ordenada, que se obtiene en este experimento como resultado de la selección de atributos, se distinga en cada factor de posicionamiento su pertenencia a la primera o segunda página que conforman las instancias ("A" o "B"). Es decir, en la lista se muestran ordenados 48 atributos que se corresponden con los 24 factores de posicionamiento tratados en la experimentación.

Los resultados de este experimento se muestran en la Figura III-11.

average merit	**average rank**	**attribute**
10023.135 +-105.565	1.5 +- 0.67	27 percentage_kw_body_B
9900.588 +-237.045	1.7 +- 0.46	3 percentage_kw_body_A
9433.309 +-71.428	3.4 +- 0.49	22 position_first_kw_A
9431.414 +-304.55	3.4 +- 0.92	46 position_first_kw_B
7119.583 +-376.754	6.4 +- 1.56	36 percentage_kw_linkedpages_B
6785.051 +-37.184	7 +- 1.41	32 percentage_kw_p_B
6970.158 +-375.84	7.5 +- 2.01	12 percentage_kw_linkedpages_A
6778.889 +-54.691	7.6 +- 1.36	8 percentage_kw_p_A
6779.433 +-53.109	7.9 +- 1.58	28 percentage_kw_links_B
6752.223 +-50.931	8.6 +- 1.36	4 percentage_kw_links_A
5766.214 +-57.461	11.5 +- 0.5	14 percentage_links_broken_A
5789.628 +-60.714	11.5 +- 0.5	38 percentage_links_broken_B
4482.167 +-39.551	13.4 +- 0.49	13 percentage_kw_incoming_links_A
4488.999 +-56.457	13.6 +- 0.49	37 percentage_kw_incoming_links_B
3497.75 +-42.007	15.5 +- 0.5	25 percentage_kw_url_B
3494.775 +-31.433	15.5 +- 0.5	1 percentage_kw_url_A
3073.751 +-29.894	17.4 +- 0.49	34 percentage_kw_metakw_B
3052.641 +-58.214	17.6 +- 0.49	10 percentage_kw_metakw_A
2584.107 +-38.189	19.4 +- 0.49	35 percentage_kw_metadescription_B
2563.992 +-72.459	19.8 +- 0.87	11 percentage_kw_metadescription_A
2408.83 +-57.955	21.3 +- 0.64	33 percentage_kw_alt_B
2418.362 +-32.406	21.5 +- 0.5	9 percentage_kw_alt_A
2060.364 +-36.038	23.5 +- 0.5	42 number_links_incoming_B
2060.385 +-37.284	23.5 +- 0.5	18 number_links_incoming_A
1880.493 +-31.666	25.5 +- 0.5	6 percentage_kw_h2_A
1887.168 +-22.269	25.5 +- 0.5	30 percentage_kw_h2_B
1471.089 +-82.879	28.1 +- 1.37	26 percentage_kw_title_B
1459.187 +-13.738	28.4 +- 0.8	5 percentage_kw_h1_A
1448.203 +-76.863	28.5 +- 1.28	2 percentage_kw_title_A
1449.757 +-33.592	29 +- 0.63	29 percentage_kw_h1_B
1219.736 +-14.902	31.5 +- 0.5	47 page_rank_B
1219.712 +-12.86	31.5 +- 0.5	23 page_rank_A
878.014 +-16.65	33.4 +- 0.49	31 percentage_kw_h3_B
877.232 +- 7.802	33.6 +- 0.49	7 percentage_kw_h3_A
339.037 +- 7.234	35.5 +- 0.5	20 number_links_classified_1_A
339.337 +-16.754	35.5 +- 0.5	44 number_links_classified_1_B
191.794 +- 3.354	37.3 +- 0.46	43 number_links_classified_0_B
191.851 +- 4.942	37.7 +- 0.46	19 number_links_classified_0_A

Figura III-11: Ranking de atributos de Google

3.3.3.2 Experimento 2: Validación de la selección de atributos de posicionamiento (Google)

La finalidad de este experimento es la aportación de evidencias sobre la validez de la metodología seguida en la selección de atributos.

Se van a seguir los pasos expuesto en el apartado correspondiente del desarrollo (3.3.2.2) a partir del ranking de factores de posicionamiento más importantes para Google (Figura III-11) obtenido en el experimento anterior.

Distinguir en el ranking de atributos los factores más importantes

La selección final de los atributos más significativos para Google se ha efectuado de forma manual. Para ello se han elegido los primeros factores de posicionamiento del ranking de la Figura III-11. El punto de corte se ha estableciendo en el primer factor con dudosa o escasa participación en el posicionamiento de Google (según evidencias literarias). En este caso el factor de corte ha sido el porcentaje de palabras clave en la meta descripción (Sullivan, 2002).

En la siguiente tabla tras eliminar las duplicidades de atributos se muestran los factores de posicionamiento que se han elegido como los más influyentes en el algoritmo de ordenación de Google.

Factores más relevantes (Google)
% of keywords in the body
Position of the first keyword
% of linked pages containing the keywords
% of keywords in the p sections
% of keywords in the links
% of broken links
% of incoming links containing the keywords
% of keywords in the URL
% of keywords in the metakw section

Tabla III-19: Factores de posicionamiento SEO más relevantes en Google

Construir con los atributos seleccionados clasificadores binarios

Las instancias utilizadas en la creación de los clasificadores binarios, a partir de consultas sobre Google, en el apartado 3.2.3 son filtradas según los atributos seleccionados (dos por cada factor seleccionado).

Una vez que las instancias se corresponden con los factores elegidos se construyen modelos binarios siguiendo la metodología empleada en la generación de los modelos binarios con todos los atributos.

En total se han obtenido seis modelos (uno por cada algoritmo de clasificación utilizado en la experimentación).

Comparación de la eficacia de los clasificadores

A los modelos creados a partir de los atributos seleccionados se les aplica evaluación cruzada (también con diez divisiones estratificadas).

En la siguiente tabla se muestra para cada algoritmo los resultados de los porcentajes de acierto de instancias correctamente clasificadas. Se comparan con los resultados obtenidos al experimentar con el conjunto total de atributos (columna del % de acierto de la Tabla III-20).

	Porcentaje de acierto	
Clasificador	**Todos los atributos**	**Atributos seleccionados**
C4.5	83,36	82,65
PART	87,20	85,31
Bagging C4.5	88,35	86,71
Bagging PART	92,01	91,12
Boosting C4.5	88,30	87,01
Boosting PART	90,46	89,30

Tabla III-20: Influencia de la selección de atributos en los porcentajes de acierto

3.3.3.3 Experimento 3: Repercusión de la selección de atributos en las estimaciones de relevancia documental (Google)

En este experimento se construyen modelos de estimación definitivos a partir de los modelos binarios generados con el conjunto reducido de atributos. La finalidad es

comparar las medidas de éxito de sus pronósticos de posicionamiento con las que se obtuvieron al intervenir todos los factores, y así ver la repercusión de la selección de atributos para las estimaciones de relevancia documental en Google.

También se construyen modelos de estimación eliminando progresivamente en cada experimento los datos de entrada correspondientes a una de las consultas que se realizaron en Google.

Los modelos se alimentaran exclusivamente con datos procedentes de la consulta "*information analysis*", otros se corresponderán con las consultas "*information analysis*" y "*genetic algorithms*", etc. Los resultados podrían aportar evidencias de que la importancia de los factores en el posicionamiento depende del buscador, pero no de las consultas que se realizan.

Los resultados obtenidos para los algoritmos C4.5 y PART con los diferentes conjuntos de entrada se presentan en la Tabla III-21 y la Tabla III-22 respectivamente:

C4.5				
Nº set de resultados	**% acierto clasificador**	**Error medio**	**Desviación estándar**	**Máximo error**
1	86,52	4,88	6,67	50
2	85,62	3,76	4,87	52
3	83,40	5,14	6,24	52
4	82,65	4,64	5,16	49

Tabla III-21: Resultados C4.5 con los atributos seleccionados

PART				
Nº set de resultados	**% acierto clasificador**	**Error medio**	**Desviación estándar**	**Máximo error**
1	88,35	4,15	6,79	51
2	88,60	4,08	4,87	47
3	86,08	4,86	5,43	47
4	85,31	4,77	5,48	47

Tabla III-22: Resultados PART con los atributos seleccionados

Los resultados obtenidos para los algoritmos de *Bagging C4.5* y *Bagging PART* se presentan en la Tabla III-23 y Tabla III-24:

Bagging-C4.5				
Nº set de resultados	**% acierto clasificador**	**Error medio**	**Desviación estándar**	**Máximo error**
1	88,64	2,99	6,41	50
2	88,75	2,48	4,18	51
3	86,94	2,91	4,99	50
4	86,71	2,84	4,53	50

Tabla III-23: Resultados Bagging- C4.5 con los atributos seleccionados

Bagging-PART				
Nº set de resultados	**% acierto clasificador**	**Error medio**	**Desviación estándar**	**Máximo error**
1	92,35	2,56	6,59	52
2	93,01	2,00	4,13	53
3	91,02	2,58	4,91	51
4	91,12	2,45	4,61	51

Tabla III-24: Resultados Bagging-PART con los atributos seleccionados

Y finalmente, los resultados obtenidos para los algoritmos *Boosting C4.5* y *Boosting PART* se muestran en la Tabla III-25 y Tabla III-26.

Boosting-C4.5				
Nº set de resultados	**% acierto clasificador**	**Error medio**	**Desviación estándar**	**Máximo error**
1	89,73	2,24	6,56	51
2	89,71	1,70	3,88	51
3	87,68	1,67	4,76	51
4	87,01	1,50	4,10	52

Tabla III-25: Resultados Boosting-C4.5 con los atributos seleccionados

Boosting-PART				
Nº set de resultados	**% acierto clasificador**	**Error medio**	**Desviación estándar**	**Máximo error**
1	90,79	2,37	6,13	47
2	91,25	1,69	3,89	51
3	89,29	1,71	4,70	51
4	89,30	1,48	4,03	50

Tabla III-26: Resultados Boosting-PART con los atributos seleccionados

Los resultados obtenidos a partir de los cuatro conjuntos de entrada (los correspondientes a todas las instancias, última fila de las tablas) se comparan con los de los experimentos en los que actuaron todos los atributos (Tabla III-10). Las siguientes gráficas (una por cada medida del éxito en la estimación de la relevancia documental) muestran la comparación de los resultados.

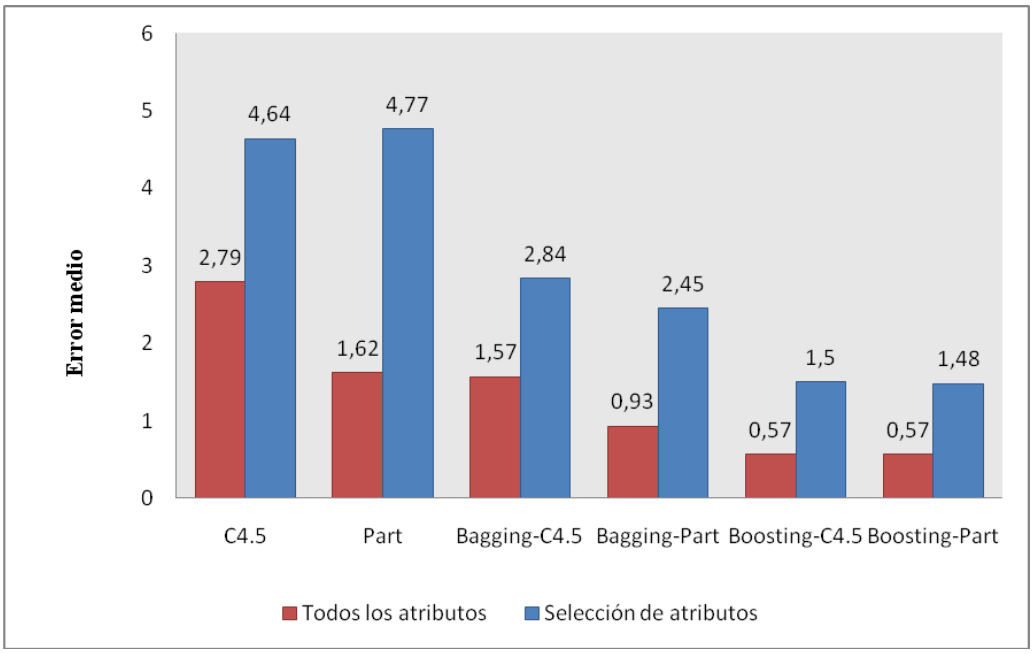


Figura III-12: Influencia de la selección de atributos en los errores medios

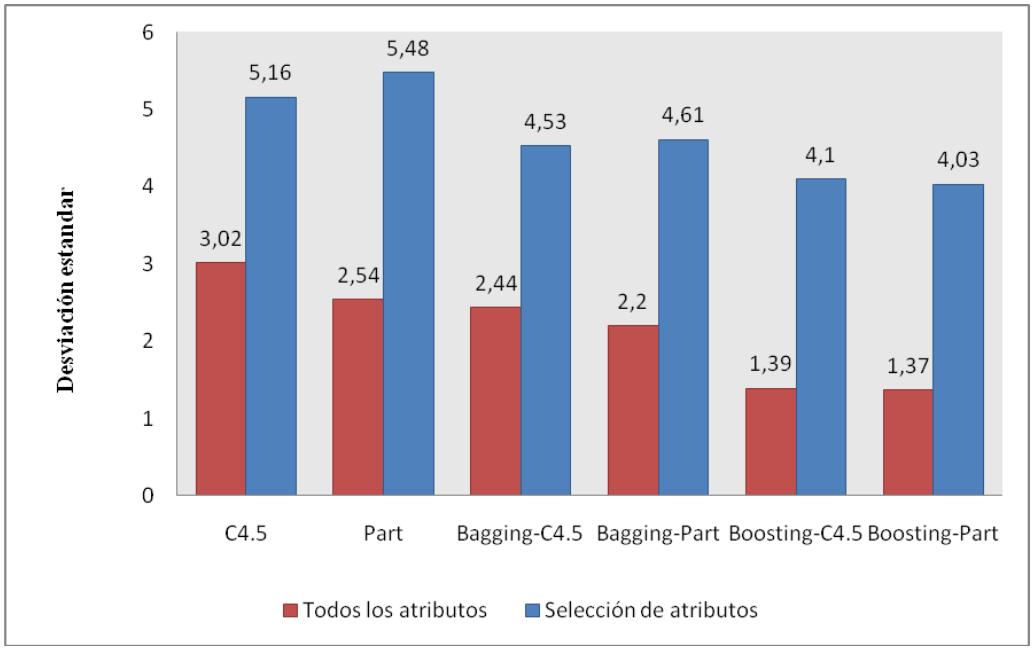


Figura III-13: Influencia de la selección de atributos en las desviaciones típicas

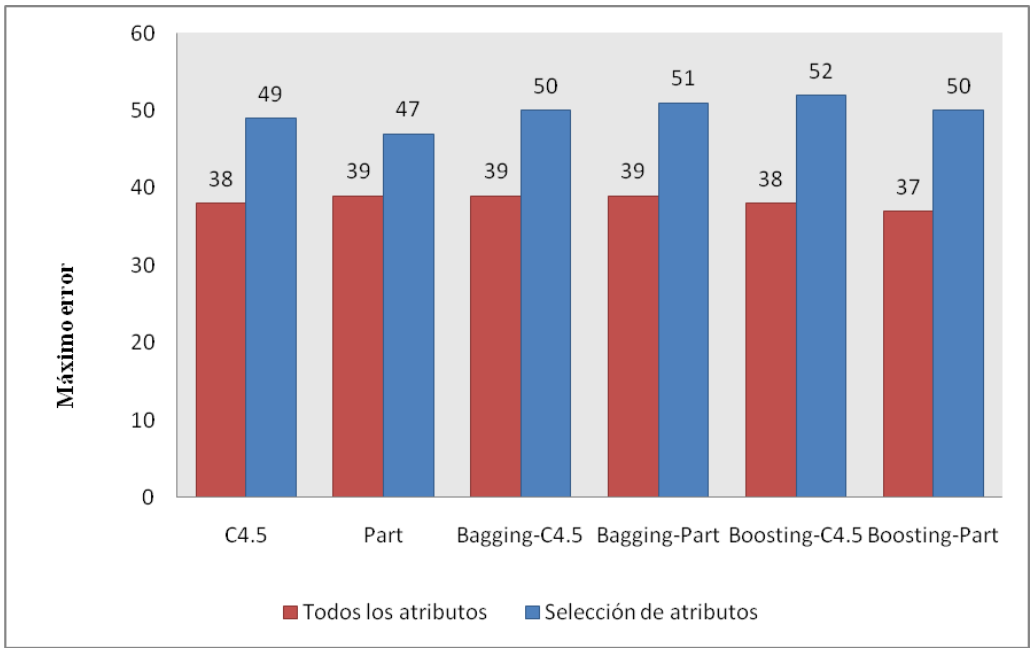


Figura III-14: Influencia de la selección de atributos en los errores máximos

A continuación se dispone una tabla resumen (Tabla III-27) con la comparativa de los resultados de los algoritmos aplicados a Google. La tabla muestra el análisis comparativo de los algoritmos C 4.5, Part, *Bagging* C 4.5, *Bagging* PART, *Boosting* .C 4.5, *Boosting* PART dependiendo de si ha utilizado todo el conjunto de atributos o sólo una selección de los mismos. En resumen, los algoritmos *Boosting* son los más adecuados en relación al error medio cometido.

Algoritmo	Set de atributos	% acierto clasificador	Error medio	Desviación estándar	Máximo error
C 4.5	Todos	83,36	2,79	3,02	38
	Selección	82,65	4,64	5,16	49
PART	Todos	87,20	1,62	2,54	39
	Selección	85,31	4,77	5,48	47
Bagging C4.5	Todos	88,35	1,57	2,44	39
	Selección	86,71	2,84	4,53	50
Bagging PART	Todos	92,01	0,93	2,2	39
	Selección	91,12	2,45	4,61	51
Boosting J48	Todos	88,30	0,57	1,39	38
	Selección	87,01	1,50	4,10	52
Boosting PART	Todos	90,46	0,57	1,37	37
	Selección	89,30	1,48	4,03	50

Tabla III-27: Comparativa de resultados de los algoritmos aplicados a Google

3.3.3.4 Experimento 4: Factores más significativos en las estrategias de posicionamiento de los motores de búsqueda Yahoo Search y MSN

Para determinar los factores más significativos en las estrategias de posicionamiento de los motores de búsqueda de Yahoo Search y MSN, se plantea un experimento. En este experimento se vuelve a aplicar los métodos ChiSquaredAttributeEval y Ranquer, tal como se procedió en el primer experimento, pero esta vez para obtener los factores más significativos de los algoritmos de ordenación de los buscadores Yahoo Search y MSN. La reducción de atributos también se llevara a cabo sobre el conjunto de factores de posicionamiento considerados para la experimentación Tabla III-6.

Las instancias de entrada son las mismas que se utilizaron respectivamente en la creación de los modelos binarios para el Yahoo Search y MSN en el experimento 2 del apartado 3.2.3. Se muestran los nueve mejores factores de posicionamiento para cada buscador y se acompañan de los nueve seleccionados para Google en el experimento 2 de esta sección (Tabla III-19) con el fin de favorecer comparaciones analíticas.

	Motor de búsqueda		
Nº	**Google**	**Yahoo**	**MSN**
1	*% of keywords in the body*	*Position of the first keyword*	*% of keywords in the body*
2	*Position of the first keyword*	*% of keywords in the body*	*Position of the first keyword*
3	*% of linked pages containing the keywords*	*% of linked pages containing the keywords*	*% of keywords in the p sections*
4	*% of keywords in the p sections*	*% of keywords in the links*	*% of linked pages containing the keywords*
5	*% of keywords in the links*	*% of keywords in the p sections*	*% of keywords in the links*
6	*% of broken links*	*% of broken links*	*% of broken links*
7	*% of incoming links containing the keywords*	*% of keywords in the URL*	*% of keywords in the metakw section*
8	*% of keywords in the URL*	*% of incoming links containing the keywords*	*% of keywords in the URL*
9	*% of keywords in the metakw section*	*% of keywords in the metakw section*	*% of keywords in the ALT sections*

Tabla III-28: Factores más significativos para cada motor de búsqueda

3.3.4 Discusión y conclusiones

El objetivo afrontado en el bloque anterior ha buscado obtener de forma automática el ranking de variables de posicionamiento que conforman el algoritmo de ordenación de un motor de búsqueda. El conocimiento de las variables de posicionamiento más influyentes permite mejorar la visibilidad de las páginas web aplicando técnicas de optimización web. Una de las conclusiones obtenidas es que la mayoría de los factores están relacionados con la presencia de palabras clave en diferentes secciones de las páginas web. Por el contrario factores populares como el PageRank no han obtenido una relevancia especialmente significativa.

Capítulo IV: Discusión y Conclusiones

En este capítulo se recoge la discusión y conclusiones de esta investigación. Se ha estructurado en diferentes apartados que se corresponden con los bloques en los que se ha organizado este trabajo, junto con un extracto de las principales conclusiones.

Como se indicó en el apartado 2.4 los trabajos relacionados con esta investigación no encajan totalmente ni con el planteamiento ni con los objetivos. Los que guardan una mayor afinidad comparan listas ordenadas de resultados para evaluar el método empleado en la predicción de ranking web. Sin embargo, en la optimización de una página web el éxito de una campaña de promoción se predice calculando los puestos que escalaría entre su competencia, y no por el grado de similitud entre dos rankings de resultados (el anterior y el posterior a la optimización). Por tanto, al no tener sentido emplear las mismas medidas en la evaluación, algunos de los resultados obtenidos en este trabajo no son comparables con los de otras investigaciones.

4.1 Discusión sobre la identificación de los factores de posicionamiento web en herramientas SEO

Como primer paso se han identificado los factores más influyentes en la relevancia de la documentación web a través de las funcionalidades más extendidas entre las herramientas SEO. Por este motivo se han descrito y analizado las funcionalidades ofrecidas por este tipo de herramientas. La permanente competencia en el mercado hace que perduren aquellas herramientas SEO que son realmente útiles en la optimización de páginas web. En resumen, los factores que evalúan deben corresponderse con los considerados por los buscadores web, en la mayor media posible, para tener opciones de conservar o incrementar su negocio.

Se han identificado 40 funcionalidades de posicionamiento. Este resultado es producto del análisis 39 herramientas SEO independientes (existen por sí mismas) de las cuales 21 son herramientas principales (no están recogidas o integradas en otras). En la Tabla IV-1 se muestran las funcionalidades de posicionamiento SEO presentes en al menos un tercio de las principales herramientas SEO.

Factores SEO más relevantes
Búsqueda palabras clave usuarios
Ranking palabras clave
Popularidad enlaces
Rastreo tráfico
Páginas indexadas
Comparación páginas primeras posiciones
Ranking tráfico
PageRank

Tabla IV-1: Factores de posicionamiento SEO más relevantes

Como se puede apreciar por el orden de las funcionalidades, los factores de posicionamiento más influyentes en el posicionamiento web están relacionados con el contenido temático de la página, seguidos de los factores de popularidad (enlaces y visitas).

También se refleja que la posición del PageRank confirma la hipótesis expuesta en (López, 2009) de que el PageRank afecta al posicionamiento de una manera más bien débil.

Por otro lado, la Tabla III-2 muestra que las herramientas más completas contemplan alrededor de 20 funcionalidades SEO. Por lo que parece posible la optimización de la relevancia web cubriendo la mitad de las funcionalidades estudiadas.

La identificación de los factores utilizados en la optimización de recursos web, su importancia atendiendo a las herramientas que los analizan y la cantidad de factores que consideran las herramientas más completas constituyen la base en la que se apoya el siguiente bloque de trabajo.

4.2 Discusión sobre la estimación de la relevancia documental asignada por los buscadores

El tema de estudio de este bloque de investigación se ha centrado en las asignaciones de relevancia que los motores de búsqueda efectúan sobre los resultados de las consultas web. Entendiendo que la relevancia de una página web se manifiesta por la posición que ocupa entre los demás resultados de la consulta, es decir, por su visibilidad. Dado que los buscadores son el medio para recuperar documentación web y que su existencia depende de apreciaciones de utilidad por parte de los usuarios, los algoritmos de posicionamiento deben ordenar los resultados de forma acorde al sentido de relevancia humano. Por tanto, los resultados obtenidos sobre la relevancia en los motores de búsqueda nos proporcionan también un reflejo de la percepción humana en cuanto a la importancia temática documental.

El enfoque metodológico propuesto se ha dirigido a la construcción automática de estimadores de la relevancia documental. Los modelos de estimación se han generado por aprendizaje del comportamiento de los buscadores web, es decir, a partir de los resultados de las consultas se han construido emuladores de los algoritmos de posicionamiento de los motores de búsqueda. De esta forma se puede predecir la importancia de un documento respecto de la temática de cierta consulta. Además, los estimadores creados permiten entender más fácilmente el funcionamiento de los buscadores y sus políticas de posicionamiento.

En general, en la creación de modelos, los resultados han sido muy satisfactorios destacando especialmente los modelos de estimación construidos a partir de algoritmos de inducción de reglas de conjuntos de clasificadores (técnicas *Boosting y Bagging*).

El algoritmo *Bagging* PART ha generado los modelos binarios más certeros en los tres buscadores con los que se ha experimentado. Dicho de otro modo, sus modelos son los que tienen mayor porcentaje de acierto, más del 90% en todos los casos, al decidir entre dos páginas cual es la más relevante.

Sin embargo, los modelos de estimación definitivos basados en los modelos binarios que se han construido por técnicas *Boosting* han alcanzado mejores medidas de éxito (media y desviación típica del error). Esta paradoja se explica por la adopción de un mecanismo de compensación de errores en el cálculo de las posiciones estimadas, tal como se explicó en la experimentación.

A la vista de los resultados se deduce que los modelos obtenidos a partir de algoritmos *Boosting* se comportan de forma muy simétrica compensando los errores de sobreestimación con los de subestimación.

Indicar que la decisión de elegir esta forma de estimar por compensación de errores, frente a otras propuestas para la eliminación de inconsistencias en la relación de orden, se tomó a partir de experimentaciones previas en las que se mostro como la mejor alternativa.

En la siguiente Tabla IV-2 se muestran los mejores resultados obtenidos por modelos de estimación con el algoritmo *Boosting* C4.5 sobre los tres buscadores (Google, Yahoo Search y MSN).

Es también importante considerar que para los buscadores Google y Yahoo Search se obtienen resultados similares con los modelos correspondientes a los algoritmos *Boosting PART* y *Boosting C4.5*.

En los resultados se aprecia que la aplicación del algoritmo *Boosting C4.5* se puede considerar como la mejor opción genérica en la construcción de modelos predictivos de la relevancia documental web.

Motor de búsqueda	Algoritmo	Error medio	Desviación estándar	Máximo error
Google	**Boosting PART**	0,57	1,37	37
Yahoo Search	**Boosting C4.5**	1,04	1,77	31
MSN	**Boosting C4.5**	1,04	1,08	33

Tabla IV-2: Resultados de los mejores modelos de estimación

Los mejores modelos de estimación por buscador presentan errores medios de aproximadamente una posición (inferior a una posición para Google) y unas desviaciones típicas relativamente pequeñas. Se demuestra, por tanto, la capacidad de la metodología propuesta para obtener modelos de estimación de la relevancia de comportamiento muy similar a los buscadores web. En otras palabras, la metodología tiene un carácter genérico de aplicación que le permite adaptarse a diferentes algoritmos de posicionamiento con el fin de obtener emuladores de sus comportamientos. Esta capacidad de aplicación con independencia del algoritmo de posicionamiento, además de ser útil para el estudio de las políticas de ordenación en diferentes motores de búsqueda, es conveniente para mantener

la utilidad de la metodología ante los ajustes y modificaciones que realizan los buscadores en sus algoritmos.

Queda por analizar en los dos primeros experimentos, los resultados del máximo error cometido. Estos son de más de 30 posiciones incluso en los mejores estimadores. No obstante, observando los valores de los errores y las medias y desviaciones típicas de los mismos se concluye que se trata de datos atípicos cuya frecuencia de aparición ronda el 0,05%. Tras analizar el sentido de las estimaciones para estas páginas se pone de manifiesto que estos errores se cometen por sobrestimación por lo que la explicación más probable tendría que ver con políticas de penalizaciones de los buscadores. Por ejemplo, los modelos construidos operan con la densidad de las palabras claves, pero por el momento son incapaces de advertir si están enmascaradas con el color de fondo u ocultas tras una imagen. Es decir, no detectan algunas políticas de posicionamiento fraudulentas que son castigadas por los buscadores.

Aunque el impacto de estos errores es mínimo por su infrecuencia sería interesante como trabajo futuro un estudio de estas situaciones. Destacar también el excelente comportamiento en el resto de predicciones, ya que a pesar de lo abultado de los errores máximos, los valores de media y desviación típica del error son muy bajos.

En el tercer experimento como se puede ver en la Figura III-7: Media del error por posición se confirma también que los errores máximos se deben a sobrestimación. Además se aprecia la uniformidad en la distribución del error medio entre las posiciones. Dicho de otra manera, la exactitud en la predicción de relevancia de un documento web es independiente de la posición real que ocupa la página.

La discusión de si los modelos generados por unas consultas determinadas son aplicables a consultas diferentes queda resuelta a la vista de los resultados del cuarto experimento. Los resultados obtenidos tras experimentar con consultas diferentes a las utilizadas en la generación de modelos son tan negativos que se puede concluir que los modelos son de aplicación específica a las consultas concretas con las que se construyeron.

Es posible que agregando los datos procedentes de muchas más consultas, a la fase de entrenamiento de los modelos, se obtuviese una aproximación más general de los protocolos de ordenación de los buscadores web. Sin embargo, las dificultades temporales para capturar características de páginas de forma automática, junto al mayor tiempo a

invertir en la creación de los modelos, complican que se pueda emular los algoritmos de posicionamiento antes de que hayan sido modificados por los buscadores.

En cuanto a la eficiencia de los algoritmos (experimento 5), se puede observar en la Figura III-8: Comparativa de tiempos de ejecución de los algoritmos que los algoritmos basados en C4.5 siguen una evolución casi lineal en el tiempo de ejecución, mientras que para los relacionados con el algoritmo PART el orden de complejidad es mayor. Además, los algoritmos base son más rápidos que sus respectivos métodos de conjuntos de clasificadores *Bagging* y estos a su vez mejoran a los *Boosting*, aunque estas diferencias parecen poco significativas.

Las medidas de tiempo se han tomado en horas, pero lo importante son las aportaciones cualitativas y no las cuantitativas ya que estas últimas dependerán de las máquinas y del paralelismo con que se ejecuten los algoritmos.

A la vista de los resultados el algoritmo que muestra un mejor rendimiento es el *Boosting* C4.5 ya que además de unos buenos resultados predictivos presenta un tiempo de ejecución de orden lineal. El último experimento demuestra su estabilidad incluso ante consultas que realizan frecuentemente los usuarios, y que por tanto, se presume una alta competitividad en el posicionamiento de sus resultados.

Finalmente, el penúltimo experimento pone de manifiesto que la configuración de los algoritmos adoptada a partir de experimentos preliminares se ha confirmado como idónea. En todos los algoritmos se alcanzan los mejores valores para al menos 3 de las 4 variables de éxito cuando el número de iteraciones es 25. Además, los resultados no mejoran a los obtenidos para el valor predeterminado de confianza del 25%.

4.3 Discusión sobre la determinación automática de la influencia de cada factor en los algoritmos de ordenación de los motores de búsqueda

En este bloque se ha propuesto una metodología para obtener de forma automática el ranking de variables de posicionamiento, según su importancia en el algoritmo de ordenación de un determinado motor de búsqueda.

Esta metodología permite por tanto conocer los factores que con mayor ponderación influyen en la visibilidad de las páginas web. En consecuencia, ayuda a dirigir los esfuerzos en pro de una buena optimización web.

Las ordenaciones de factores de posicionamiento se han obtenido a partir de instancias que representan las características de dos web indicando la más relevante. Por tanto, como se puede observar en la Figura IV-1, los atributos de posicionamiento aparecen en el ranking por partida doble. En este sentido la ordenación de los atributos se muestra consecuente, agrupando, casi con total fidelidad, las parejas de atributos correspondientes a un mismo factor de posicionamiento.

Los resultados obtenidos en el experimento 2 para la evaluación de la metodología (Tabla III-20) indican que la reducción del número de atributos no afecta significativamente a la eficacia de los clasificadores C4.5 y *Bagging* PART desde el punto de vista estadístico (diferencias menores que el 1% en el porcentaje de acierto de individuos correctamente clasificados) (Hall y Holmes, 2002). Cabe destacar que el algoritmo *Bagging* PART es el que mejor se ha comportado, antes y después de la selección, en la construcción de modelos binarios. Además, el algoritmo C4.5 es de los más comúnmente usados en las pruebas de verificación de efectividad de selección de atributos debido a su enfoque explícito hacia los atributos relevantes (Hall y Holmes, 2002; Morales y Sierra, 2006). Para el resto de algoritmos las diferencias en los porcentajes de acierto inducidas por la selección nunca han alcanzado el 2%.

Se debe poner especial atención en que el objetivo perseguido por la selección no ha sido en ningún caso mejorar el resultado de los clasificadores (para ese cometido hubiesen sido más apropiados métodos de selección de atributos *Wrappers* (Morales, 2007)), si no demostrar que el método aplicado proporciona información correcta sobre la importancia individual de cada factor de posicionamiento.

Por las razones expuestas, se puede valorar de forma positiva la metodología aplicada para inferir la influencia que ejercen en el posicionamiento de un motor de búsqueda cada uno de los factores. Asimismo, seleccionar exactamente los nueve mejores factores de Google (Tabla III-19) para esta evaluación, de forma manual, ha resultado acertado debido al conocimiento sobre el problema (Witten y Frank, 2005).

Por otra parte, las medidas de éxito de los estimadores de la relevancia generados con los nueve mejores factores del ranking de Google son peores que las obtenidas por los estimadores en los que participaban todos los factores (Figuras IV-8, IV-9 y IV-2). Esto era de esperar, ya que aunque los nueve factores seleccionados son los más influyentes, el peso conjunto de los quince factores omitidos conlleva una perdida en la calidad de las

estimaciones. No obstante se obtienen algunos modelos similares a los conseguidos con los métodos de conjuntos de clasificadores Boosting, con un error medio entorno a 1,5 posiciones.

También parece confirmarse la hipótesis lógica de que la importancia de los factores es propia de los buscadores y no dependen de las consultas. Los resultados mostrados de la Tabla III-21 y la Tabla III-26 son muy similares para los distintos conjuntos de entrada procedentes de las consultas.

Comparando ahora los mejores factores de posicionamiento obtenidos para los tres buscadores (Tabla III-28), se aprecia que salvo el orden casi coinciden a la perfección[27].

Por lo tanto, es aconsejable centrar la atención en estas variables al diseñar una página web, sabiendo que los motores de búsqueda, en general, las consideran más a la hora de asignar visibilidad.

Es notorio señalar que la mayoría de los factores están relacionados con la presencia de palabras clave en varias secciones de las páginas web. Por lo tanto, no es conveniente intentar optimizar una web para un número excesivo de consultas.

El único de los parámetros que no está relacionado con la aparición de palabras clave es el porcentaje de enlaces rotos. Su importancia radica en la penalización que sufren las páginas que los contienen. Los buscadores consideran que estas páginas han tenido un diseño descuidado o un deficiente mantenimiento, en cualquier caso un indicio de mala calidad.

Es destacable que variables tan populares como el PageRank no tenga demasiada relevancia en el posicionamiento, ni siquiera para Google, como se comprobó también en el estudio de las variables SEO. Según López (2009) esto sucede para el buscador Google desde 2008 por modificaciones en sus políticas. Otras sin embargo si eran esperables, como las relacionadas con las palabras clave asociadas a los enlaces entrantes, cuyo poder ha quedado demostrado por los famosos casos de *bombing*.

[27] Tan sólo hay una discrepancia (en la columna relativa a MSN aparece el factor: "% *of keywords in the ALT sections"* en lugar de *"% of incoming links containing the keywords"*.

4.4 Extracto de las principales conclusiones

En este apartado se recogen las principales conclusiones a las que se ha llegado en esta investigación. Se pueden consultar con más detalle en los apartados de discusión y conclusiones correspondientes a cada uno de los tres bloques en los que se ha dividido la investigación.

En relación con las 40 funcionalidades identificadas en el estudio de las principales herramientas SEO, se observa que la que tienen más presencia en estas herramientas son las que tienen que ver con el análisis de palabras clave, seguidas de aquellas que están orientadas a un aumento de la popularidad (análisis de enlaces y tráfico). De lo que se deduce que los factores de posicionamiento asociados a estas funcionalidades son los más influyentes en la visibilidad de de los documentos web. No siendo el PageRank uno de los más destacados.

Después de aplicar técnicas de selección de atributos sobre las características de posicionamiento se ha llegado a otra conclusión: los motores de búsqueda Google, Yahoo Search y MSN coinciden en los factores de posicionamiento que más influyen en sus algoritmos de ordenación. Al igual que se ha observado en el estudio de las herramientas SEO, los factores más destacados guardan relación con las palabras clave y con la popularidad web, coincidiendo también en la discreta influencia del PageRank. Esto indica, como se suponía, que las herramientas SEO reflejan ciertas características de los buscadores, siendo útiles para realizar estudios indirectos de los mismos.

En cuanto a la predicción de la relevancia documental web, los modelos de estimación han resultado muy precisos al emular la asignación de relevancia temática de los buscadores (Google, Yahoo Search y MSN). En particular los modelos binarios obtenidos con el algoritmo *Bagging* PART superan el 90% de acierto al seleccionar entre dos páginas la más relevante. Mientras que los modelos definitivos construidos a partir del algoritmo *Boosting* C4.5 cometen errores medios cercanos a una posición (incluso inferiores: 0,57 posiciones para Google, 0,84 si son consultas frecuentes) en las estimaciones de las posiciones de páginas web entre sus competidoras. Por tanto, la metodología propuesta tiene capacidad para adaptarse a distintos algoritmos de ordenación pudiendo emular sus comportamientos.

Es destacable que los errores medios producidos en los pronósticos de los modelos se distribuyen uniformemente entre todas las posiciones, por tanto el éxito en la predicción de la posición de una web es independiente de su posición real.

En la eficiencia de los algoritmos utilizados en la construcción de los modelos, destacan aquellos que utilizan como base el algoritmo C4.5, con un coste computacional de orden lineal. Esto convierte al algoritmo *Boosting* C4.5 en el más adecuado en la construcción de modelos, ya que destaca simultáneamente en eficacia y eficiencia.

Respecto a la generalidad de los modelos, los experimentos demuestran que se pueden incluir resultados de distintas consultas en la generación de un modelo y obtener muy buenos resultados en predicciones relacionadas con esas consultas. Sin embargo, si se desea hacer predicciones asociadas a consultas diferentes con ese mismo modelo no se obtienen buenos resultados. Es decir, para que un modelo estime de forma adecuada la relevancia relativa a la temática de una consulta, se debe incluir en su generación instancias de los resultados de dicha consulta.

Hay que enfatizar que el avance más estable realizado en el campo del posicionamiento web es la metodología desarrollada, capaz de adaptarse de forma automática al dinamismo de la relevancia en la Web. Es decir, aunque el análisis de los resultados obtenidos es interesante para descubrir tendencias y rutinas a la hora de posicionar documentos, su utilidad puede ser pasajera. Lo importante es la capacidad demostrada para realizar este tipo de análisis en cualquier momento, incluso cuando se hayan modificado las reglas de juego por motores de búsqueda, webmasters o usuarios.

El proceso metodológico propuesto en la presente investigación para la estimación de la relevancia documental web solventa el problema de determinar con antelación los costes y mejoras que se proponen en la optimización web. Tal como se ha descrito en esta tesis se pueden generar modelos predictivos que estimen la posición de un documento web frente a los de su competencia a partir de los valores de sus atributos. Si se modifican estos atributos atendiendo a las mejoras que se podrían practicar sobre ese documento se determinará la posición que alcanzaría de realizarse el proceso de optimización. Este método es previo a realizar el proceso de optimización, evaluando el grado de mejora que se va a conseguir. A su vez, se pueden determinar las características a modificar que tendría mayor impacto en la visibilidad del documento.

Este enfoque da lugar a lo que podría llamarse herramientas SEO de nueva generación y ha sido registrado en una patente internacional, en el campo de métodos de optimización web, por la Universidad Carlos III de Madrid (2009).

Capítulo V: Trabajos Futuros

Los resultados obtenidos en esta tesis han sido óptimos por lo que no queda un amplio margen para mejorarlos significativamente, sin embargo quedan pendientes aspectos por donde seguir desarrollando esta investigación.

Uno de estos aspectos es el idioma utilizado en la experimentación. Se ha elegido la lengua inglesa en las consultas realizadas en los motores de búsqueda por ser el idioma más internacional, científica y tecnológicamente hablando. En consecuencia, en las herramientas multilingües o con versiones para varios idiomas, es para el primero que se desarrollan, ajustan y funcionan de forma estable. Por otro lado, se presume que los algoritmos de posicionamiento de los buscadores web actúan de la misma forma al ordenar documentos que comparten el mismo idioma, ya que el tratamiento de términos es similar salvo por detalles específicos como el cálculo de términos flexionados por lematización (Martí y Llisterri, 2002). No obstante, parece interesante, como trabajo futuro, experimentar en otras lenguas para confirmar que la relevancia web es independiente del idioma.

Otra línea a seguir es la construcción de clasificadores binarios para la estimación de la relevancia documental en la web. Estos clasificadores deciden entre dos documentos web cual es el más pertinente para determinada consulta por emulación de un motor de búsqueda. Los clasificadores se han generado por técnicas de inducción de reglas por sus ventajas frente a otros algoritmos, sin embargo, no se descarta que otros métodos de aprendizaje sean igualmente útiles y que por tanto sean el centro de futuras investigaciones.

Las herramientas SEO que se han analizado limitan sus optimizaciones a páginas web asociadas a código HTML. Están diseñadas para extraer información a partir de las

etiquetas que estructuran este código y algunas disponen de editores para modificarlo. En los análisis de web que se han realizado en esta tesis se ha dado un paso más al incorporar también los archivos con formato PDF. A pesar de ello, han quedado sin tratar otros formatos, con porcentajes exiguos de aparición en la Web, en los que cabe investigar como asignan los buscadores su relevancia.

Respecto a los factores de posicionamiento, la restricción de los recursos económicos ha impedido la participación de todos ellos en la experimentación. A la vista de los resultados estos factores omitidos parece que no son demasiado influyentes en la visibilidad de los documentos, o bien están en correlación con otros que han asumido su representación. No obstante, una mayor inversión permitiría, mediante herramientas especializadas de pago, ampliar la captura automática de los valores de los factores de posicionamiento no tratados.

Por último, se podría analizar las causas de los errores máximos cometidos en las predicciones de relevancia de páginas web. Estos errores, con un porcentaje de ocurrencia ínfimo, probablemente se deban a políticas de penalizaciones de los buscadores (como ya se comento) o volviendo al punto anterior, a factores que no se han podido considerar en los experimentos.

Referencias

Agosti, M., & Melucci, M. (2000). Information Retrieval on the Web. *Lecture Notes in Computer Science*, 1980. Springer Verlag, Berlin Heidelberg, 242- 285.

Antoniou, G., & Harmelen, F. van. (2004). A Semantic Web Primer. *London: The MIT Press*.

Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., & Raghavan, S. (2001). Searching the Web. *ACM Transactions on Internet Technology*, 1 (1), 2-43.

Babiak, U. (1999). Effekive Suche im Internet. O'Reilly Verlag, Köln.

Baeza-Yates, R., & Ribeiro-neto, B. (1999). Modern information retrieval. *New York: ACM Press* (Vol. 24). New York; Harlow; Madrid: ACM Press; Addison-Wesley.

Baeza-Yates, R., Saint-Jean, F., & Castillo, C. (2002). Web Structure, Dynamics and Page Quality. *Lecture Notes in Computer Science*, 2476. Laender, A., Oliveira, A. (Eds.), Springer Verlag, Berlin Heidelberg, 117-130.

Baeza-Yates, R. (2003). Information Retrieval in the Web Beyond Current Search Engines. *International Journal of Approximate Reasoning*, 34, 97-104.

Bar-Ilan, J. (2005). Comparing rankings of search results on the web. *Information Processing and Management*, 41, 1511-1519.

Bharat, K., & Broder, A. (1998). A technique for measuring the relative size and overlap of public Web search engines. *Computer Networks and ISDN Systems*, 30 (1–7), 379–388.

Bookstein, A. (1983). Outline of a general probabilistic retrieval model, *Journal of Documentation 39*, 2, 63-72.

Bradford SC (1948). Documentation. London: Crosby Lockwood & Sons.

Bradman, O., Cho, J., Garcia-Molina, H., & Shivakumar, N. (2000). Crawler-Friendly Web Servers. *ACM SIGMETRICS Performance Evaluation Review*, 28 (2), 9-14.

Breiman L., Bagging predictors. (1996) *Machine Learning*, 24(2):123–140.

Brin, S. & Page, L., (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks*, 30, 107-117.

Callon, M., Courtial, J.P. & Penan, H., (1995). Cienciometría. La medición de la actividad científica: de la bibliometría a la vigilancia tecnológica. *Gijón: Trea.*

Callon M., Courtial J.P., Penan H. (1993) La scientométrie. Paris, PUF, « Que Sais-je ? », 2727, 126.

Cao, Y., Xu, J., Liu, T.-Y., Li, H., Huang, Y., & Hon, H.-W. (2006). Adapting ranking SVM to document retrieval. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval*, ACM, Seattle, Washington, USA.

Castro, L.R. de, & Timmis, J. (2003). Artificial immune systems as a novel soft computing paradigm. *Soft Computing* 7 (8), 526–544.

Cendrowska J. (1987). PRISM: An algorithm for inducing modular rules. *International Journal of Man-Machine Studies*, 7 (4), 349-370.

Chang, G., Healey, M., McHugh, J., & Wang, J. (2001). Mining the World Wide Web: An Information Search Approach. Kluwer Academic Publishers, Boston.

Chau, M., & Chen, H. (2003). Personalized and Focused Web Spiders. *Web Intelligence, Zhong, N. et al. (Eds.)*. Springer Verlag, Berlin Heidelberg, 197-216.

Chaumier, J., & Dejean, M. (1990). L'indexation documentaire: de l'analyse conceptuelle humaine a l'analyse automatique morpho-syntaxique. *Documentalist, 27 (6)*, 275-279.

Chen, Y., Gan, Q., & Suel T. (2002). I/O-Efficient Techniques for Computing Pagerank. *Proceedings of the eleventh international conference on information and knowledge management*, 549-557.

Chien, S., Dwork, C., Kumar, R., Simon, D. R., & Sivakumar, D. (2003). Link evolution: Analysis and algorithms. *Internet Mathematics*, *1*(3), 277-304.

Chignell, M. H., Gwizdka, J., & Bodner, R. C. 1999. Discriminating meta-search: A framework for evaluation. *Information Processing and Management*, 35, 337–362.

Choi, O., Yoon, S., Oh, M., & Han, S. (2003). Semantic Web Search Model for Information Retrieval of the Semantic Data. *Lecture Notes in Computer Science*, 2713. Chung, C. et al. (Eds.), Springer Verlag, Berlin Heidelberg, 588-593.

Chowdhury, G., & Chowdhury, S. (2001). Searching CD-ROM and Online Information Sources. Library Association Publishing, London.

Clark, C. V. (1957). America`s psicologists: a survey of a growing profession. *America Psychological Association, Washington, DC.*

Clark, P., & Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning*, *3*, 261-283.

Clay, B. (2009). Search Engine Relationship Chart. http://www.bruceclay.com/ searchenginerelationshipchart.htm, [Consulta: 06/2010].

Cleveland, D. B., & Cleveland, A. D. (1990). Introduction to Indexing and Abstracting. *Colorado: Libraries Unlimited*.

Craswell, N. Hawking, D., & Robertson, S. (2001). Effective Site Finding using Link Anchor Information. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 250-257.

Codina, L. (2004). Posicionamiento Web: Conceptos y Ciclo de Vida [online]. *Hipertext.net*, *núm2* (ISSN 1695-5498). http://www.hipertext.net, [Consulta: 05/2009]. ISSN 1695-5498.

Cohen, J. (1995). Highlights: Language and Domain-Independent Automatic Indexing Terms for Abstracting. *Journal of the American Society for Information Science, 46(3)*.

Cole, F. J., & Eales, N. B. (1917). The history of comparative anatomy. Part I: a statistical analysis of the literature. *Science Progress*, *11*, 578-596.

Cole, J. R., & Cole, S. (1971). Measuring the quality of sociological research: problems in the use of the Science Citation Index. American Sociologist, 6, 23-29.

Cowie, J., & Wilks, Y. (2000). Information Extration. *En DALE, R. (ed). Handbook of Natural Language Processing. New York: Marcel Dekker*, 241-260.

Cummins, R., O'Riordan, C. (2005). Evolving General Term-Weighting Schemes for Information Retrieval: Tests on Larger Collections. Artif. Intell. Rev. 24(3-4), 277-299.

Dietterich T. G. (2000). Ensemble methods in machine learning. In J. Kittler and F. Roli, editors, Multiple Classifiers Systems: first international workshop; proceedings /MCS 2000, volume 1857 of Lecture Notes in Computer Science, pages 1–15, Cagliari, Italy, June 2000. Springer.

Dietterich T. G. (1997). Machine-learning research: four current directions. AI Magazine, 18 (4), 97–136.

Diligenti, M., Gori, M., & Maggini, M. (2002). Web Page Scoring Systems for Horizontal and Vertical Search. *Proceedings of the eleventh international conference on World Wide Web*, 508-516.

Diligenti, M., Fellow, M. G., & Maggini, M. (2004). A unified probabilistic framework for web page scoring systems. *IEEE transactions on knowledge and data engineering*, *16*(1), 4-16.

Ding, W., & Marchionini, G. (1998). A comparative study of Web search service performance. *In Proceedings of the annual conference of the American Society for Information Science*. 136–142.

Dogpile.com. (2007). Different engines, different results web searchers: Not alwaysfinding what they're looking for online. Available at: www.infospaceinc.com/onlineprod/Overlap-DifferentEnginesDifferentResults.pdf

Dvorak, J. C. (2004). Search Engine Mania. *The Mad Rush to Develop New Ways of Finding Info Online, PC Magazine*, http://www.pcmag.com/article2/0,2817,1555066,00.asp, [Consulta: 06/2010].

Eastman, C., & Jansen, B. (2003). Coverage, Relevance and Ranking: The Impact of Query Operators on Web Search Engine Results. *ACM Transactions on Information Systems*, 21 (4), 383-411.

Egghe, L., & Rousseau, R. (1990). Introduction to informetrics: quantitative methods in library documentation and information science. *Elservier, Ámsterdam*.

Egghe, L., & Rousseau, R. (2005). Classical retrieval and overlap measures satisfy the requirements for rankings based on a Lorenz curve. *Information Processing and Management*, 42 (1), 106–120.

Elliott, J., & Eckstein, V. (2002). Java Swing, O'Reilly & Associates; 2nd edition, November 1, 2002.

Frakes, W., & Baeza-Yates, R. (1992). Information Retrieval: Data structures and Algorithms. *Upper Saddle River: Prentice-Hall.*

Fensel, D., Bussler, C., Ding, Y., Kartseva, V., Klein, M., Korotkiy, M., Omelayenko, B., & Siebes, R. (2002). Semantic Web Application Areas. *Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems. Stockholm, Sweden.*

Ferber, R. (2003). Information Retrieval - Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web. *dpunkt.verlag*. Heidelberg.

Freund, Y., Iyer, R., Schapire, R.E., & Singer, Y. (2003). An efficient boosting algorithm for combining preferences, *Journal of Machine Learning Research* 4, 933–969.

Freund, J., Miller, I., & Miller, M. (2000) Estadística matemática con aplicaciones. Pearson –Prentice-Hall. Sexta edición.

Freund, Y., & Schapire, R. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. In Springer-Verlag, editor, Proceedings of the Second European Conference on Computational Learning Theory, pages 23–37.

Freund, Y., & Schapire, R. (1996) Experiment with a new boosting algorithm. In M. Kaufmann, editor, Proceedings of the Thirteenth International Conference on Machine Learning, 148–156.

Fuhr, N. (2000). Models in Information Retrieval. *Lecture Notes in Computer Science*, 1980. Agosti, M. et al. (Eds.), Springer Verlag, Berlin Heidelberg, 21-50.

Giles, C., Petinot, Y., Teregowda, P., Han, H., Lawrence, S., Rangaswamy, A., & Pal, N. (2003). eBizSearch: A Niche Search Engine for e-Business. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, 413-414.

Gilyarevsky, R., Uzilevsky, G., & Moudrov, E. (1997). An automatic statistical classification of different types of journals. *International Forum on Information and Documentation, 22*(3), 24-35.

Glöggler, M. (2003). Suchmaschinen im Internet. Springer Verlag, Berlin Heidelberg.

Gordon, M. (1998). Probabilistic and genetic algorithms in document retrieval. *Communications of the ACM, 31 (10)*, 1208-18.

Hall, M. A. (1998). Correlation-based Feature Selection for Machine Learning. *PhD Thesis. University of Waikato, Department of Computer Science, Hamilton, New Zealand*.

Hall, M. A., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, *11 (1)*, 10-18.

Hall, M. A., & Holmes, G. (2002). Benchmarking Attribute Selection Techniques for Data Mining. Technical Report 00/10. *University of Waikato, Department of Computer Science, Hamilton, New Zealand, Julio*. Hamilton, New Zealand. Retrieved from http://www.cs.waikato.ac.nz/~ml/publications/2000/00MH-GHBenchmarking.pdf. [Consulta: 02/2009].

Harman, D. (1994). Automatic Indexing. Challenges in indexing electronic text and images (ASIS monograph series). *Medford (New Jersey): Learned Information*, 247-264.

Hatzivassiloglou, V., Klavans, V., & Eskin, E. (1999). Detecting text similarity over short passages: exploring linguistic feature combinations via machine learning. *Proceedings of the EMLNP/VLC99 Joint SIGDAT* Conference on Empirical Methods in NLP and Very Large Corpora. University of Maryland, College Park, Maryland.

Hawking, D., Craswell,N., Thistlewaite, P., & Harman, D. (1999). Results and Challenges in Web Search Evaluation. Proceedings of the WWW Conference, pp. 244-252.

Henzinger, M. (2000). Link Analysis in Web Information Retrieval. *IEEE Data Engineering Bulletin*, 23 (3), 3-8.

Henzinger, M. (2000a). Web Information Retrieval-an Algorithmic Perspectiva. *Proceedings of the 8th Annual European Symposium on Algorithms*, 1-8.

Henzinger, M. (2001). Hyperlink Analysis for the Web. *IEEE Internet Computing*, 1 (5), 45-50.

Hodges, J., Yie, S., Raighart, R., & Boggess, L. (1996). An automated systems that assists in the generation of document indexer. *Natural Language Engineering, 2 (2)*, 137-160.

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning, 11*, 63-91.

Hong, J., Mozetic, I., & Michalski, R. S. (1986). AQ15: Incremental learning of attribute based descriptions from examples, the method and user's guide. *In Report ISG 85-5UIUCDCS-F*. Department of Computer Science, University of Illinois at Urbana-Champaign.

Ingwersen, P. (2000). Users in Context. Lecture Notes in Computer Science, 1980. Agosti, M. et al. (Eds.), Springer Verlag, Berlin Heidelberg. 157-178.

Introna, L., & Nissenbaum, H. (2000). Shaping the Web: Why the Politics of Search Engines Matters. *The Information Society*, 16, 169-185.

Jeh, G., & Widom, J. (2003). Scaling Personalized Web Search. *Proceedings of the twelfth international conference on World Wide Web*, 271-279.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining, ACM*.

John, R., & Mooney, G. (2001). Fuzzy User Modeling for Information Retrieval on the World Wide Web. *Knowledge and Information Systems*, 3, 81-95, Springer Verlag, London.

Jux2.com. (2004). Search engines are more different than people think. http://jux2.com/stats.php, [Consulta: 06/2010].

Kan, M., & Thi, H. O. (2005). Fast webpage classification using URL features. *In Proc. CIKM, Bremen, Germany*.

Katz, B., Felshin, S., Yuret, D., Ibrahim, A., Lin, J., Marton, G., McFarland, A. J. & Temelkuran, B. (2002). Omnibase: Uniform Access to Heterogeneous Data for Question Answering. *Proceedings of the seventh International Workshop on Applications of Natural Language to Information Systems*.

Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. *ACMSIAM Symposium on Discrete Algorithms (SODA)*, *46*(5), 604-632. http//wwwcscornelledu/home/kleinber/authpdf. [Consulta: 02/2009].

Kleinberg, J. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46 (5), 604-632.

Kline, V. (2002). Missing links: the quest for better search tools. *Online Information Review*, 26 (4), 252-255.

Knudsen, J., & Niemeyer, P. (2005). Learning Java, 3rd Edition. *O'Reilly Media*, Inc., Sebastopol, CA, USA.

Kobayashi, M., & Takeda, K. (2000). Information Retrieval on the Web. *ACM Computing Surveys*, 32 (2), 144-173.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial In*, *97*(1-2), 273-324.

Kwok, C., Etzioni, O., & Weld, D. S. (2001). Scaling Question Answering to the Web. ACM Transactions on Information Systems, 19 (3), 242-262.

Lawani, S. M. (1977). Citation Analysis and the Quality of Scientific Productivity. *BioScience*, *27*(1), 26-31. http://www.jstor.org/stable/1297790 [Consulta: 09/2009].

Lawrence, S., & Giles, C. (1999). Searching the web: General and scientific information access. *IEEE Communications*, *37*(1), 116-122.

Lin, J., Quan, D., Sinha, V., Bakshi, K., Huynh, D., Katz, B., & Karger, D.R. (2003). The Role of Context in Question Answering Systems. Proceedings of the 2003 Conference on Human Factors in Computing Systems.

Liu, H., & Setiono, R., (1996). A probabilistic approach to feature selection - A filter solution. In *13th International Conference on Machine Learning*. Morgan Kauffman. 319-327.

López, M., & Práctico. (2009). Posicionamiento En Buscadores. *Marketing Online Taller/Curso Práctico*. www.consultoresvalencia.com. [Consulta 09/2009].

Lorenzo J. (2002). Selección de Atributos en Aprendizaje Automático basado en la Teoría de la Información. PhD thesis. Faculty of Computer Science, Univ. of Las Palmas. Gran Canaria.

Losee, R. M. (1996). Text windows and phrases differing by discipline, location in document, and syntactic structure. *Information Processing & Management*, *32*(6), 747-767.

Lovins, J.B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics* 11, 22–31.

Ludwig, M. (2003). Breaking Trough the Invisible Web. *Net connect, Winter 2003*, 8-10.

Luhn, H.P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development 1958, 2(2)*, 159-195.

Machill, M., Neuberger, C., & Schindler, F. (2002). Transparenz im Netz, Funktionen und Defizite von Internet-Suchmaschinen. Verlag Bertelsmann Stiftung, Gütersloh.

Machill, M., Neuberger, C., Schweiger, W., & Wirth, W. (2003). Wegweiser im Netz: Qualität und Nutzung von Suchmaschinen. Wegweiser im Netz, Machill, M., Welp, C. (Eds.), Verlag Bertelsmann Stiftung, Gütersloh.

Major, J., & Mangano, J. (1995). Selecting among rules induced from a hurricane database. *Journal of Intelligent Information Systems, 4*, 39-52.

Marckini, F. (2001). Search Engine Positioning. *Plano, Texas: Wordware Publishing Inc., Isbn 155622804x.*

Marendy, P. (2001). A review of world wide web searching techniques. http:// citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.12.2055, [Consulta 06/2010].

Markoff, J. (2004). Google Planning to Roll Out E-Mail Service. *New York Times.* http:// www.nytimes.com/2004/03/31/technology/31CND-GOOGLE.html?ex=1081569600& en=642bb461d24a7ec5&ei=5070#, [Consulta 06/20010].

Martí, M. A., & Llisterri, J. (2002). Tratamiento del lenguaje natural. *Barcelona: Universitat de Barcelona*, 207.

Martin, B. R., & Irvine, J. (1983). Assessing basic research: Some partial indicators of scientific progress in radio astronomy. *Research Policy*, 12(2), 61-90.

McGuigan, G. (2003). Invisible Business Information: the Selection on Invisible Web Sites in Construction Subject Pages for Business. *Collection Building,* 22 (2), 68-74.

Moens, M. (2000). Automatic Indexing and Abstracting of Documents. Kluwer Academic Publishers, Boston.

Moldovan, D., & Surdeanu, M. (2003). On the Role of Information Retrieval and Information Extraction in Question Answering Systems. *Lecture Notes in Artificial Intelligence* 2700. Springer Verlag, Berlin Heidelberg, 129-147.

Molina, L. C., Belanche, L., & Nebot, Á. (2002). Evaluación de Algoritmos de Selección de Atributos. *CCIA*. http://www.lsi.upc.es/~lcmolina/SC/html/paper/ccia02-fs.pdf. [Consulta 06/2010].

Morales, E. (2007). Cursos. Aprendizaje. Selección de atributos. http://ccc.inaoep.mx/~emorales/Cursos/Aprendizaje2/seleccion.pdf. [Consulta 06/2010].

Morales, E., & Sierra Araujo, B. (2006). *Aprendizaje Automático: conceptos básicos y avanzados. Aspectos prácticos utilizando el software WEKA*. Pearson. Prentice Hall.

Morato, J. (1999). Análisis de relaciones cienciométricas y lingüisticas en un entorno automatizado. *PhD Thesis. Universidad CARLOS III DE MADRID, Departamento de Biblioteconomía y Documentación.*

Morato, J., Sánchez Cuadrado, S., & Valiente, M. (2005). Análisis de las estrategias de posicionamiento en relación a la relevancia documental. *El profesional de la Información*, 18.

Morato, J., Sánchez Cuadrado, S., & Marrero, M. (2009). Sistemas Avanzados de Recuperación de Información. Retrieved from OpenCourseWare - Universidad Carlos III de Madrid. Web site: http://ocw.uc3m.es/informatica/sistemas-avanzados-de-recuperacion-de-informacion.

Moreno, V. (2005). Interacción entre medidas de popularidad en el posicionamiento web. *El profesional de la información*, *14*, 100-107.

Morgan, J., & Kilgour, A. (1996). Personalising On-line Information Retrieval Support with a Genetic algorithm. *PolyModel, 16: Applications of artificial intelligence*, 142-149.

Moss, J. (2001). Guidelines and information on internet search engines and website promotion. http://www.cral.ac.uk/guidelines/search/searcheng.htm. [Consulta 06/2010].

Musilek, P., Lau, A., Reformat, M., & Wyard-Scott, L. (2006). Immune programming, *Information Sciences* 176 (8), 972–1002.

Nicholson, S. (2000). Raising reliability of Web search tool research through replication and chaos theory. *Journal of the American Society for Information Science*, 51 (8), 724–729.

Paice, C. D. (1990). Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management, 26 (1)*, 171-186.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. Technical report. *Computer Science Dept., Stanford Univ*.

Paulson, J. (2003). Fast Facts About Froogle. SitePoint. http://articles.sitepoint.com /article/fast-facts-froogle, [Consulta: 06/2010].

Pontigo, J., & Lancaster, F. W. (1986). Qualitative aspects of the Bradford distribution. *Scientometrics*, *9*(1-2), 59-70.

Pretto, L. (2002). A Theoretical Analysis of Google's PageRank. *Lecture Notes in Computer Science*, 2476. Leander, A., Oliveira, A. (Eds.), Springer Verlag, Berlin Heidelberg, 131-144.

Price, L., & Thelwall, M. (2005). The Clustering Power of Low Frequency Words in Academic Webs. *Journal of The American Society For Information Science*, 56(8), 883–888.

Quinlan, J. R. (1986). Induction of Decision Trees (ID3 algorithm). *Machine Learning*, *1*(1), 81-106.

Quinlan, J. R. (1993). C4.5: Programs for Machine Learnirig. *Morgan Kaufmann, CA*.

Richardson, M., & Domingos, P. (2002). The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank. *Advances in Neural Information Processing Systems*, 14. MIT Press, Cambridge, 1441-1448.

Richardson, M., & Domingos, P. (2004). Combining Link and Content Information in Web Search. *M. Levene and A. Poulovassilis (eds.), Web Dynamics*, 179-193. New York: Springer.

Rijsbergen, C. van. (1979). Information retrieval. *Butterworth*.

Robertson, S. (2000). Evaluation in Information Retrieval, in: Lecture Notes in Computer Science, 1980. Agosti, M. et al. (Eds.). Springer Verlag, Berlin Heidelberg. 81-92

Robertson, S. (2002). Comparing the Performance of Adaptive Filtering and Ranked Output Systems. Information Retrieval, 5, 257-268

Rodríguez-Miñón, P. (1982). Estadística (aplicada a la Biología): curso de nivelación A.T.S. Universidad Nacional de Educación a Distancia, UNED. Spain. ISBN: 84-362-1444-7.

Roebuck, M. (2000). Search Engine Placement and Ranking. *PA: Infinity Publishing.Com*.

Rossi, G., Schwabe, D., & Guimaraes, R. M. (2001). Designing Personalized Web Applications. *Proceedings of the Tenth International WWW Conference*, Hong Kong, 275-284.

Rumbaugh, J., Jacobson, I., & Booch, G. (1998). The unified modeling language reference manual. *Massachusetts: Addison-Wesley.*

Salton, G. (1989). Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. *Reading, MA: Addison-Wesley*.

Salton, G., & McGill, M.J. (1983). An introduction to modern information retrieval. *McGraw-Hill*.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning, 5*, 197-227.

Schimkat, R. Küchlin, W., & Nestel, F. (2002). Living Hypertext – Web Retrieval Techniques. *Lecture Notes in Computer Science*, 2346. Unger, H. et al. (Eds.), Springer Verlag, Berlin Heidelberg, 1-14.

Schwartz, C. (2001). Sorting Out the Web. *Ablex Publishing. London.*

Schultz, C.K., Luhn, H.P. (1968). Pioneer of information science. *Spartan Books.* New York, NY.

Seglen, P. O. (1996). Quantification of scientific article contents. *Scientometrics, 35*(3), 335-366.

Sherman, C. (2003). Help Test the Wondir Search Engine. http://searchenginewatch.com /2208541. [Consulta: 06/2010].

Silverstein, C., Henzinger, M., & Marais, H. (1998). Analysis of a Very Large Web Search Engine Query Log. *Digital SRC Technical Note*, 1998-014.

Smart, J. C. (1983). Perceived quality and citation rates of education journals. *Research in Higher Education*, *19*(2), 175-182.

Spyns, P., Pretorius A. J., & Reinberger, M. L. (2004). Evaluating DOGMA-lexons generated automatically from a text corpus. Belgica: STAR Lab Technical Report. Retrieved from http://www.starlab.vub.ac.be/website/node/297.

Spyns, P., & Reinberger, M. L. (2005). Evaluating ontology triples generated automatically from texts. In A. Gomez-Perez y Euzenat, J.,(eds.), *Proceedings of the second European Conference on the Semantic Web, LNCS 3532. Springer*, 563 - 577

Stallman, R. M., McGrath, R., & Smith, P. D. (2004) GNU make: a program for directed recompilation. Version 3.81. *Free Software Foundation.*

Sterling, G. (2004). Local Search: The Hybrid Future. http://searchenginewatch.com/ searchday/article.php/3296721, [Consulta: 06/2010].

Sullivan, D. (2002). Death of a meta tag. http://searchenginewatch.com/2165061, [Consulta 06/2010].

Sullivan, D. (2003). Major Search Engines and Directories. http://www. searchenginewatch.com/links/article.php/2156221, [Consulta 06/2010].

Sullivan, D. (2003a). Search Privacy at Google & Other Search Engines, http:// searchenginewatch.com/2189531, [Consulta: 06/2010].

Sullivan, D. (2004). Google Launches Gmail, a Free Email Service. http:// searchenginewatch.com/_subscribers/articles/article.php/3334251, [Consulta 06/2010].

Thompson, D. (2002) The influence of metatags on web-based search retrieval, ranking and relevancy, Submitted in partial fulfillment of the requirements for the degree of Executive Master of Electronic Commerce at Dalhousie University Halifax, Nova Scotia, April, 2002.

Thompson, B. (2003). Is Google too Powerful? *BBC News.* http://news.bbc.co.uk/2/hi/ technology/2786761.stm, [Consulta: 06/2010].

Trotman, A. (2005). Choosing document structure weights. Inf. Process. Manage. 41(2), 243-264.

Türker, D. (2004). The optimal design of a seach engine from an Agency Theory perspective. *Working paper, Institute for Broadcasting Economics – University of Koeln.*

Universidad Carlos III de Madrid. (2009). Procedimiento y sistema de estimación de la posición de un recurso. *Moreno, V., Morato, J., & Sánchez-Cuadrado, S.* España. Patente Internacional. PCT/ES2009/070517. 2009-11-20

Vazirgiannis M., Drosos D., Senellart P., and Vlachou A. (2008) Web Page Rank Prediction with Markov Models,WWW poster, Beijing, China, 2008.

Velasco, M., Lloréns, J., & Martínez, V. (1997). Generación Automática de Representaciones de Dominios. II Jornadas en Ingeniería de Software. JIS97. San Sebastián. España.

Wang, S., Ma , J., & He, Q. (2010). An immune programming-based ranking function discovery approach for effective information retrieval. *Expert Systems with Applications: An International Journal* 37 (8), 5863-5871. ISSN: 0957-4174.

Weiss, S. M., & Indurkhya, N. (1998). Predictive Data Mining. *San Francisco: Morgan Kaufmann*.

Weiss, S. M., & Kulikowski, C. A. (1991). Computer Systems That Learn: Classification and Prediction Methodsfrom Statistics, Neural Nets, Machine Learning, and Expert Systems. *San Mateo, CA: Morgan Kaufmann*.

Wensi, X., Fox, E. A., Tan, R. P., & Shu, J. (2002). Machine Learning Approach for Homepage Finding Task. *Lecture Notes in Computer Science*, 2476. Leander, A., Oliveira, A. (Eds.), Springer Verlag, Berlin Heidelberg, 145-159.

Witten, I. H., & Frank, E. (2005). *Data Mining. Practical Machine Learning Tools and Techniques, 2th Ed. Morgan Kaufmann Publishers*. Ed. Morgan Kaufmann Publishers.

Wolpert, D. (1992). Stacked generalization. *Neural Networks*, *5*, 241-259.

Xu, J., & Li, H. (2007). AdaRank: A boosting algorithm for information retrieval. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, ACM*, Amsterdam, The Netherlands.

Yang, H., King, I., & Lyu, M. R. (2005). Predictive ranking: a novel page ranking approach by estimating the Web structure. http://portal.acm.org/citation.cfm?id=1062810&dl=GUIDE&coll=GUIDE&CFID=91375299&CFTOKEN=56697782. doi: 10.1145/1062745.1062810. [Consulta 06/2010].

Zacharouli P., Titsias M., Vazirgiannis M., (2009). Web Page Rank Prediction with PCA and EM Clustering, Proceedings of the 6th International Workshop on Algorithms and Models for the Web-Graph, February 12-13, 2009, Barcelona, Spain.

Zadeh, L.A. (1965). Fuzzy sets. *Information and Control 8*, 3, 338-353.

Zahdeh, L. (2003). From Search Engines to Question-Answering Systems–The Need For New Tools. http://www.eecs.berkeley.edu/~shawnc/bisctalk_slides/LotfiTalkAbstract.pdf.

Zipf GH (1949) Human behaviour and the principle of least effort. Cambridge, Mass: Addison Wesley.

Anexo A: Producción científica asociada a esta investigación

En este apartado se incluyen algunos resultados de investigaciones relacionados con este trabajo, que se han realizado previamente. Son incluidos en esta tesis, porque han supuesto la base para este trabajo.

El primer resultado que se incluye como producción científica es una patente internacional que registra la idea central de esta tesis y cuya titularidad corresponde a la Universidad Carlos III de Madrid (2009). Se presenta como un procedimiento sistematizado que puede predecir la posición que ocuparía un recurso, entre otros ordenados que se consideran competencia, si sus características fuesen modificadas. El sistema, aplicado con miras a la promoción de un recurso, permite evaluar la rentabilidad de una campaña de optimización con anterioridad a su puesta en marca. De esta forma, en caso de no ser viable se pueden proponer otras alternativas hasta encontrar alguna, si es que existe, que sea pronosticada como adecuada por el sistema. Como puede apreciar se ha generalizado el ámbito de aplicación de las ideas contenidas en la tesis, extendiéndose más allá de la relevancia web.

Otro de los resultados fue un artículo en una revista que estudia la influencia de los factores de popularidad en el posicionamiento web (Moreno, 2005). En este trabajo se muestra las interrelaciones de los factores que miden la popularidad de un sitio web por parte de los usuarios, con los que evalúan la consideración o autoridad que otras páginas web le otorgan. Estas correspondencias entre ambos tipos de factores son utilizadas por algunos buscadores para defenderse de políticas fraudulentas como el uso de granjas de enlaces. De esta forma si un motor de búsqueda detecta en una web un tráfico escaso en relación al número de enlaces que recibe, entonces activa sus protocolos de penalización.

1

PROCEDIMIENTO Y SISTEMA DE ESTIMACIÓN DE LA POSICIÓN DE UN RECURSO

ANTECEDENTES DE LA INVENCIÓN

Campo técnico

La presente invención se refiere a un procedimiento y sistema de estimación de la posición, que un recuperador de información proporciona a un recurso con respecto a recursos de la competencia y más particularmente a la estimación de la posición que obtendrá un documento Web ante cualquier consulta realizada a través de un motor de búsqueda.

Descripción de la técnica relacionada

La búsqueda de información no se entiende sin los motores de búsqueda Web. Esto genera un alto grado de competitividad para que las páginas Web sean recuperadas en las primeras posiciones, porque los usuarios sólo suelen consultar los primeros resultados. Los usuarios consideran que los primeros resultados que devuelve un motor de búsqueda son los más relevantes para su consulta. Esto se traduce en mayor prestigio y consideración para las páginas mejor posicionadas.

Por tanto, hay un interés considerable en crear las páginas Web de tal modo que destaquen ante determinadas consultas. Los motores de búsqueda determinan la relevancia de páginas Web ante una determinada consulta usando los valores de ciertas características de la página, por ejemplo palabras clave en el enlace, palabras en el título, etc.

En la mayoría de los ocasiones se conocen cuáles son las características usadas por los motores de búsqueda, pero lo que se desconoce es cómo influyen en una ordenación por relevancia. Por tanto, se podría decir que el algoritmo de ranking de los motores de búsqueda es "no-público" o secreto.

Existen métodos para mejorar el posicionamiento de páginas Web con respecto a su competencia para un determinado motor de búsqueda y ante una determinada consulta. Se conocen tales métodos por el nombre SEO

(*Search Engine Optimization*, Optimización de Motor de Búsqueda). Una herramienta SEO analiza una página de acuerdo a una o más características y da consejos de cómo mejorarla. Por ejemplo, se puede analizar los valores de ciertas características de páginas mejor posicionadas que la página considerada. Es probable que, si se copia el valor de una característica presente en todas las páginas mejor posicionadas, haya una mejora de la posición, pero se desconoce a priori cuantas posiciones se pueden superar basándose en tales consejos. Por tanto, esas propuestas no prevén si el esfuerzo invertido resulta rentable con las mejoras resultantes, o si las políticas de optimización empleadas son adecuadas. Estas soluciones resultan insuficientes, ya que no aseguran ni pueden predecir las posiciones que las páginas Web ocuparán después de aplicar las hipotéticas mejoras de optimización.

Otros recursos, como los rankings de instituciones sufren del mismo problema que las páginas Web. Generalmente, se conocen las características o los factores que deciden el ranking, aunque el peso de cada factor suele ser secreto.

DESCRIPCIÓN DE LA INVENCIÓN

La invención tiene como objetivo proporcionar un procedimiento y un sistema que solucionen los problemas anteriormente mencionados.

Para ello, según la invención se proporciona un procedimiento y sistema según las reivindicaciones independientes. Realizaciones favorables se definen en las reivindicaciones dependientes.

Según un primer aspecto de la invención, se proporciona un procedimiento de estimación de la posición, que un recuperador de información proporciona a un recurso con respecto a los recursos de su competencia. El recuperador de información asigna la posición de los recursos usando los valores de las características de ellos (de los recursos) mediante un algoritmo de ranking. El procedimiento comprende el paso de la creación de un modelo que emula el algoritmo de ranking del recuperador de

información. Se realiza la creación del modelo a partir de los valores de al menos una de las características de los recursos de la competencia y las posiciones que el recuperador de información previamente ha otorgado a los recursos de la competencia. Se aplica el modelo para estimar la posición del recurso a posicionar con respecto a los recursos de la competencia.

Preferiblemente, los recursos son documentos Web y el recuperador de información es un motor de búsqueda, que estima la posición de los documentos Web ante una determinada consulta.

Este procedimiento permite estimar la posición que determinado motor de búsqueda (a priori, no está restringido a ninguno) otorgará a un documento Web, por ejemplo una página Web, ante determinada consulta. La estimación puede ser previa a que esa página Web sea indizada e incluso antes de que sea explorada por el rastreador.

El procedimiento también permite la estimación de ranking en el caso de una posible optimización Web de una página ya indizada. Incluso previamente a realizar la optimización, se podría saber el efecto que causaría en su posicionamiento.

Aunque la aplicación preferida del procedimiento según la invención es la predicción de la posición de recursos Web, se puede aplicar a otros recursos, por ejemplo a rankings de instituciones o la probabilidad de venta de un producto basado en su publicación en la Web.

La metodología ofrece mayor ventaja competitiva a los que lo apliquen para dilucidar el posicionamiento de un ranking basado en criterios no públicos.

Según una realización de la presente invención, el procedimiento comprende el paso de la extracción de los valores de las características de posicionamiento que se consideran relevantes tanto para el documento a posicionar, como para los documentos que representan su competencia. Se entiende por documentos que representan su competencia, los documentos que aparecen como resultados asociados a la consulta en función del recuperador de información elegido.

A partir de los datos obtenidos para los documentos de la competencia, se construye un modelo mediante técnicas de aprendizaje automático que permite discernir si un documento se posicionará mejor que otro. Se obtiene también información estadística relativa a su fiabilidad. El modelo se construye preferiblemente mediante aprendizaje automático con reglas de inducción, no obstante, también se podría crear el modelo mediante redes de neuronas, máquinas vectoriales, algoritmos genéticos, y otros sistemas orientados al aprendizaje automático.

Derivado del modelo construido se obtiene las relaciones de orden entre el documento a posicionar y cada uno de los documentos de su competencia. Se usa la información estadística para resolver posibles incoherencias en las relaciones de orden, y se determina la posición más probable para el documento.

Preferiblemente, el procedimiento según la invención se implementa mediante un programa informático.

Según un segundo aspecto adicional de la invención, se proporciona un sistema que está configurado para realizar las etapas del procedimiento escrito anteriormente.

Estos y otros aspectos de la invención resultarán evidentes a partir de y se dilucidarán con referencia a las realizaciones descritas a continuación.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

La invención, la técnica y las ventajas de sus objetivos resultarán más evidentes para los expertos con la explicación de la técnica mediante los siguientes dibujos, junto con la memoria descriptiva que los acompaña, en los que:

La figura 1 muestra un diagrama de flujo de la estimación de la posición de un recurso según una realización de la presente invención.

La figura 2 representa los atributos de una instancia de aprendizaje usada para crear el modelo, que emula el algoritmo del recuperador de información.

La figura 3 muestra un diagrama de flujo del proceso de optimización de las características de un recurso.

La figura 4 representa un sistema para implementar los pasos del diagrama de flujo de la figura 1.

En todas las figuras los mismos números de referencia se refieren a elementos iguales.

DESCRIPCIÓN DETALLADA DE LA INVENCIÓN

A la vista de las figuras reseñadas, puede describirse aquí una realización práctica de la invención.

Según dicha realización de la invención, se propone un procedimiento para la estimación de la relevancia documental de un documento para una consulta determinada respecto a los documentos de su competencia. En concreto, el procedimiento se centra en el posicionamiento de documentación Web (por ejemplo páginas Web) ante consultas concretas para cualquier motor de búsqueda.

A partir de este procedimiento, se puede determinar la posición que alcanzará un documento tras realizar un proceso de optimización. Este procedimiento es previo a realizar el proceso de optimización, evaluando el grado de mejora que se va a conseguir.

A su vez, se pueden determinar las características a modificar que tienen mayor impacto en la visibilidad del documento. Se entiende por visibilidad del documento la posición en los resultados de un motor de búsqueda online ante una consulta.

La figura 1 muestra la estimación 100 de la posición de una página a optimizar 150. El procedimiento se puede dividir en varias fases.

Primero, se fija 110 la consulta a realizar y el motor de búsqueda en el que se va a realizar la consulta, por ejemplo Google, Yahoo Search, MSN etc.. Habitualmente, los usuarios de motores de búsqueda realizan consultas expresadas por palabras clave.

Luego, se determinan las páginas de la competencia 120 de la página a optimizar. Para conseguirlo se ejecuta la consulta fijada en el motor de búsqueda elegido y se seleccionan las primeras n páginas recuperadas por el buscador escogido. Por ejemplo, se puede seleccionar los cien primeros resultados (n=100) por considerar que es un número suficiente de páginas, en el sentido de constituir un conjunto de datos significativo, y a su vez no tan amplio como para incluir páginas de mala calidad y por tanto poco competitivas. Sin embargo, el número 100 es relativo y en un principio, si existe la posibilidad, se puede incluso seleccionar todos los resultados, aunque se estima que podría dar mucho ruido en los contextos en los que la calidad de todas las páginas no haya sido evaluada con antelación a su inclusión en el repositorio.

Una vez obtenidas las páginas resultantes de una consulta se procede a la extracción 130 de los valores de sus características, es decir, los valores de una serie de atributos fijados de antemano cuyo propósito es poder discernir la calidad de cada página en función de su visibilidad (posición que le ha otorgado el motor de búsqueda entre las páginas obtenidas en esa consulta). Se trata de características habituales que utilizan los motores de búsqueda, por ejemplo: palabras clave en el enlace, palabras en el titulo, número de enlaces entrantes, la popularidad de los enlaces, etc. Se puede realizar la obtención de los valores correspondientes a las características mediante un programa que captura estos datos automáticamente.

Este último paso plantea el problema de cuáles son las características ideales para medir la relevancia documental de una página Web y cómo obtener los valores de esas características. Este problema es crucial a la hora de poder representar un documento Web para la finalidad perseguida. En la mayoría de los ocasiones se conocen cuáles son esas características y lo que se desconoce es cómo influyen en una ordenación por relevancia. Frecuentemente, los propios buscadores aportan información de las características o atributos que intervienen en el posicionamiento. Además,

foros y herramientas especializadas en optimización Web (herramientas SEO *Search Engine Optimization*, Optimización de Motor de Búsqueda) hacen públicos los parámetros que utilizan.

La creación 140 del modelo que emula al algoritmo del motor de búsqueda se realiza por aprendizaje automático a partir de los valores de las características de las páginas que constituyen la competencia, atendiendo al posicionamiento que les ha asignado el motor de búsqueda, previamente.

El objetivo del modelo es poder decidir entre dos páginas Web cual es la más relevante ante cierta consulta. Para su creación se pueden utilizar algoritmos de inteligencia artificial supervisados, en concreto algoritmos de inducción de reglas.

Este tipo de algoritmos tiene las siguientes ventajas:

- La robustez frente al ruido (debidos a errores, omisiones o insuficiencia de datos)
- Identificación de atributos irrelevantes
- Detección de la ausencia de atributos discriminantes y de vacíos de conocimiento
- Extracción de reglas fáciles de entender y de gran expresividad
- Posibilidad de reprocesar las reglas mediante el conocimiento de expertos, interpretando, modificando o aceptando reglas

Posibles algoritmos de inducción de reglas que se puede usar para la creación del modelo son C 4.5 y Rules Part. También se puede combinar clasificadores homogéneos para mejorar la precisión mediante las técnicas *Bagging* y *Boosting*, tomando como base los algoritmos anteriores. No obstante, otros tipos de algoritmos como redes de neuronas, máquinas vectoriales, algoritmos genéticos, etc. también son apropiados para la construcción del modelo.

Las técnicas de inducción de reglas reciben la información como un conjunto de casos (ejemplos de aprendizaje). Estos ejemplos se representan por un conjunto de atributos común que incluye el atributo de clase. Los valores de estos atributos distinguen unos casos de otros. A partir de los

datos de entrada generan un árbol de decisión o un conjunto de reglas que proporcionará la clasificación de los nuevos ejemplos.

En este caso, cada instancia o ejemplo de aprendizaje, independientemente del algoritmo utilizado, se construye a partir de cada pareja de páginas que se pueden formar con las n primeras páginas de una consulta. Por tanto, el número de instancias máximo es V(n,2) (variaciones de cien elementos tomados de dos en dos). Como muestra la figura 2, los atributos de cada instancia son los valores 200 de las características de la primera página más los valores 210 de las características de la segunda página más el atributo de clase 220. Este último atributo contiene la información de cuál de las dos páginas es más relevante y se obtiene a partir de las posiciones que ocupan en los resultados de la consulta.

El modelo creado puede predecir cuál es la mejor posición entre dos páginas para una determinada consulta y en función de los valores de sus características.

Ahora para saber la posición que un motor de búsqueda otorgará a una página Web 150 ante una determinada consulta se aplica 160 el modelo repetidamente para obtener la relevancia de dicha página 150 frente a cada una de las páginas 120 de la competencia. El modelo sólo requiere los valores de las características de la página 150 y por tanto, se puede estimar la relevancia de ese documento sin necesidad de que esté indexado previamente. Asimismo, se puede determinar el impacto que un documento Web sufrirá en su posicionamiento ante hipotéticas modificaciones de optimización.

A continuación se procede a la eliminación 170 de posibles inconsistencias en la relación de orden de las páginas. Como los modelos obtenidos por aprendizaje automático difícilmente son precisos en un 100%, podrían surgir inconsistencias respecto de la definición de relación de orden, ya que podrían incumplirse las propiedades de antisimétrica y transitiva. Es decir, podría ser que la página 1 estimase que es mejor que la 2, la página 2 mejor que la 3 y la página 3 mejor que la 1. Estos defectos o errores de

precisión del modelo se subsanan mediante estadística, corrigiendo aquellas predicciones cuya confianza sea más improbable, o de menor confianza, bien por estar en discordancia con respecto al resto de las predicciones, o bien corrigiendo la predicción más incoherente respecto al resto de las predicciones. Para esto se utilizan métodos estadísticos a partir de matrices de confusión, probabilidades de las decisiones, etc.

Por ejemplo, se puede optar por corregir el mínimo número de predicciones que permita establecer la relación de orden. También se puede reforzar esta idea teniendo en cuenta las probabilidades de decisión que cada regla del modelo tiene asociada.

Mediante la aplicación 160 del modelo y la eliminación 170 de posibles inconsistencias en la relación de orden de las páginas, se consigue la estimación 180 de la página 150 con respecto a las páginas de la competencia 120.

Se puede validar el procedimiento, reservando páginas de los resultados de las consultas. Es decir, se valida el procedimiento con las páginas que se han reservado y no han intervenido ni en la creación 140 del modelo ni en la fase 170 de eliminación de inconsistencias. Estas páginas permiten medir la precisión para la estimación de la posición de una página. Comparando la posición estimada y la posición real (posición en la lista de resultados facilitado por el motor de búsqueda ante la consulta) de cada una de las páginas reservadas se obtiene la media y la desviación del error cometidos respecto al número de posiciones.

El procedimiento de estimación de la posición 100 descrito anteriormente se puede usar en el proceso de optimización 300 de una página web, como muestra la figura 3. Se propone 310 una modificación de los valores de las características y antes de modificarla se procede a la estimación de la posición 100 de la página hipotéticamente modificada. Se determina 320 si la modificación es rentable por ejemplo en términos del número de posiciones que se gana con ella. La rentabilidad de la modificación de la página es un factor que depende del propietario de la

página web a optimizar. Si es rentable se procede a la realización de la optimización 330 y se finaliza 340 el proceso. Si no es rentable, se decide si se sigue 350 con el proceso de optimización. En caso afirmativo, se propone una modificación, de otro modo se finaliza el proceso. Las propuestas de alteración se pueden hacer de forma manual o siguiendo las recomendaciones obtenidas de forma automática sobre lo que diferencia a una página de las de su competencia.

Se puede implementar el procedimiento según la invención mediante un programa informático cargado en una memoria asociada 420 a un procesador 410 de un sistema informático 400, que muestra la figura 4.

Aunque la invención se ha ilustrado y descrito en detalle en los dibujos y en la descripción precedente, tal ilustración y descripción deben considerarse ilustrativas y no restrictivas; la invención no está limitada a las realizaciones dadas a conocer.

De este modo, no sólo se puede aplicar el procedimiento según la invención para la predicción de la posición de recursos Web, como por ejemplo páginas Web, sino que se puede aplicar a otros recursos, por ejemplo rankings de instituciones o probabilidad de venta de un producto basado en su publicación en la Web.

También, se puede aplicar el procedimiento a sistemas de vigilancia. Algunos sistemas de vigilancia reportan una alerta si detectan alguna modificación en el ranking de algún recurso. Es un servicio que se ofrece, porque los propietarios quieren saber si algo ha cambiado respecto a su página Web para solucionarlo, por ejemplo si la posición de su página web ha bajado mucho dentro de los resultados de una determinada consulta. El procedimiento según la invención permite estimar el esfuerzo para solucionar esa pérdida de posiciones y el motivo por el que se han perdido posiciones en el ranking.

Los expertos en la técnica pueden entender y realizar otras variaciones de las realizaciones dadas a conocer a la hora de poner en práctica la invención reivindicada, a partir de un estudio de los dibujos, la

descripción, y las reivindicaciones adjuntas. En las reivindicaciones, el término "que comprende" no excluye otros elementos o etapas, y el artículo indefinido "un" o "una" no excluye una pluralidad. Un único procesador u otra unidad pueden desempeñar las funciones de varios artículos enumerados en las reivindicaciones. El mero hecho de que se enumeren determinadas medidas en reivindicaciones dependientes mutuamente diferentes no indica que no pueda usarse ventajosamente una combinación de esas medidas. Un programa informático puede estar almacenado / distribuirse en un medio adecuado, tal como un medio de almacenamiento óptico o un medio de estado sólido suministrado junto con o como parte de otro hardware, aunque también puede distribuirse de otras formas, tales como a través de Internet u otros sistemas de telecomunicación por cable o inalámbrica. Ningún símbolo de referencia en las reivindicaciones debe interpretarse como que limita el alcance.

12

REIVINDICACIONES

1. Procedimiento de estimación de la posición en el ranking, que un recuperador de información otorga a un recurso (150) con respecto a recursos de la competencia (120), en el que el recuperador de información estima la posición de los recursos usando valores de características de ellos mediante un algoritmo de ranking, en el que el procedimiento se caracteriza por los siguientes pasos:
   - la creación (140) de un modelo que emula el algoritmo de ranking del recuperador de información a partir de los siguientes datos de entrada:
     - valores de al menos una de las características de los recursos de competencia; y
     - las posiciones que el recuperador de información previamente ha otorgado a los recursos de competencia;
   - la aplicación (150) del modelo para estimar la posición del recurso a posicionar con respecto a los recursos de competencia.-
2. Procedimiento según la reivindicación 1, caracterizado porque los recursos son documentos Web y el recuperador de información es un motor de búsqueda, que estima la posición de los documentos Web ante una determinada consulta.
3. Procedimiento según la reivindicación 1 ó 2, caracterizado porque se construye el modelo mediante inducción de reglas.
4. Procedimiento según la reivindicación 1 ó 2, caracterizado porque se construye el modelo mediante redes de neuronas, máquinas vectoriales o algoritmos genéticos u otros algoritmos de Inteligencia Artificial.
5. Procedimiento según cualquiera de las reivindicaciones anteriores, caracterizado porque se construye el modelo a partir de instancias de parejas de recursos de competencia, en el que cada instancia comprende los valores (200,210) de al menos una de las

características de ambos recursos y un atributo (220) que indica cuál de ambos recursos ha sido mejor posicionado por el recuperador de información.

6. Procedimiento según cualquiera de las reivindicaciones anteriores, caracterizado porque el modelo permite discernir cuál de una pareja de recursos se posiciona mejor.

7. Procedimiento según la reivindicación 6, caracterizado porque se aplica el modelo para discernir la posición del recurso a posicionar con respecto a cada uno de los recursos de la competencia.

8. Procedimiento según la reivindicación 7, caracterizado por el paso adicional de resolver posibles incoherencias en el orden de la posición del recurso a posicionar y las posiciones de los recursos de competencia.

9. Procedimiento según la reivindicación 8, caracterizado porque se utiliza información estadística para resolver las posibles incoherencias.

10. Un programa informático que comprende medios de código de programa informático adaptados para realizar las etapas de una cualquiera de reivindicaciones 1 a 9, cuando dicho programa se ejecuta en un ordenador.

11. Procedimiento de optimización de un recurso caracterizado porque comprende los siguientes pasos:

- proponer (310) una modificación de los valores de las características del recurso;
- antes de implementar la modificación, estimar (100) la posición del recurso hipotéticamente modificado usando el procedimiento según una cualquiera de reivindicaciones 1 a 9,
- determinar (320) si la modificación es rentable y
- si es rentable proceder a la realización de la optimización (330).

12. Sistema (400) que está configurado para realizar las etapas de una

cualquiera de reivindicaciones 1 a 9.

15

RESUMEN

Procedimiento y sistema de estimación de la posición, que un recuperador de información otorga a un recurso (150) con respecto a recursos de la competencia (120). El recuperador de información estima la posición de los recursos usando valores de características de ellos mediante un algoritmo de ranking. El procedimiento comprende el paso de la creación (140) de un modelo que emula el algoritmo de ranking del recuperador de información. Se realiza la creación del modelo a partir de los valores de al menos una de las características de los recursos de la competencia y las posiciones que el recuperador de información previamente ha otorgado a los recursos de la competencia. Se aplica (150) el modelo para estimar la posición del recurso a posicionar con respecto a los recursos de competencia.

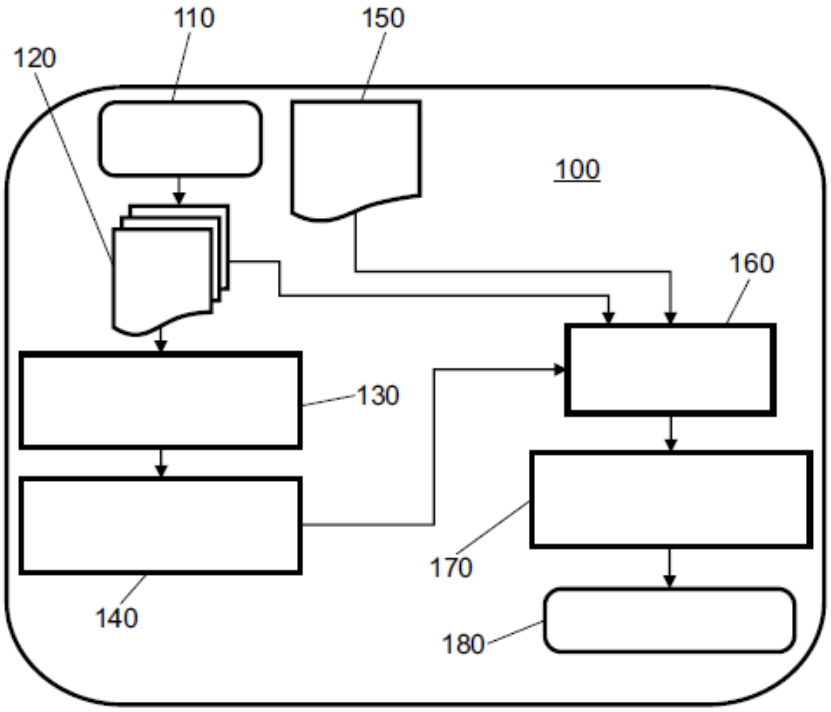

110
150
120
100
160
130
170
140
180

FIG. 1

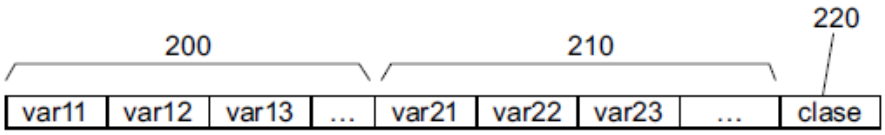

200
210
220
var11
var12
var13
...
var21
var22
var23
...
clase

FIG. 2

2/2

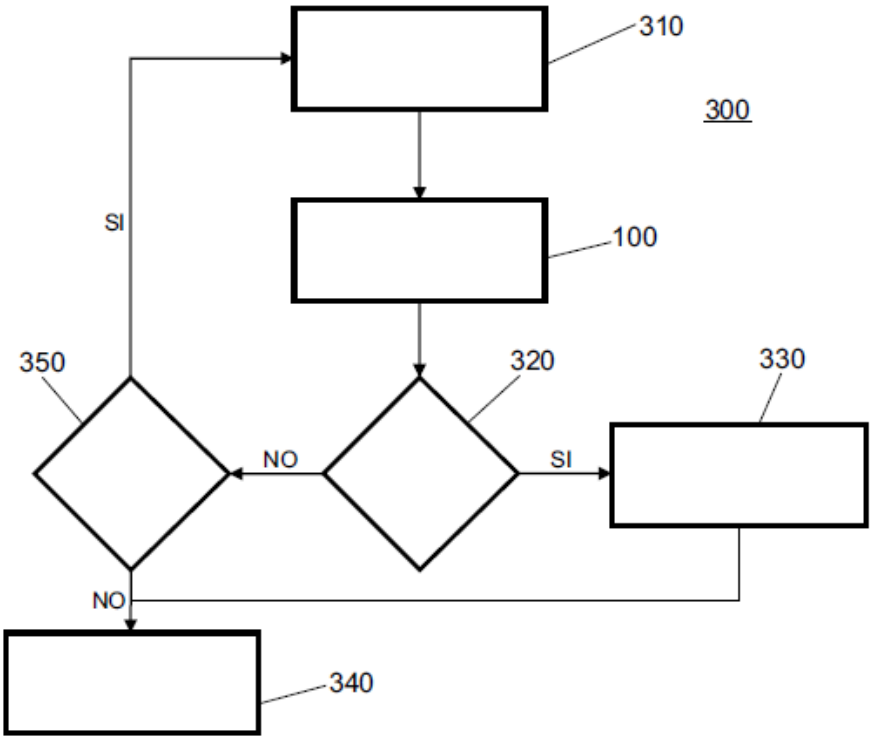


FIG. 3

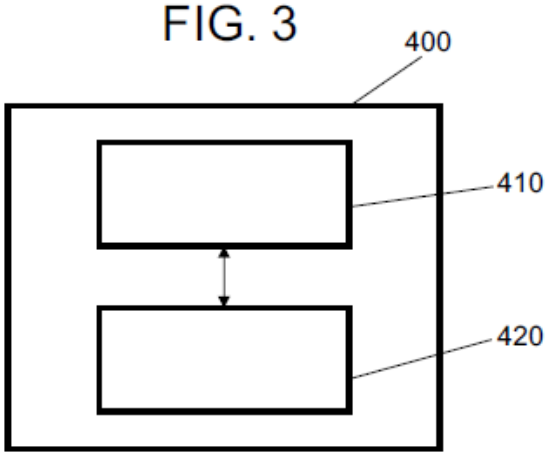


FIG. 4

Artículo

Interacción entre medidas de popularidad en el posicionamiento web

Por **Valentín Moreno Pelayo**

Resumen: *El trabajo se centra en la relación entre algunas medidas de popularidad y posicionamiento de las páginas recuperadas en los buscadores Google y MSN Search. Se ha analizado la interacción entre el número de enlaces a una página y su tráfico. Se ha obtenido como conclusión que en Google, de forma más decisiva que en MSN Search, el posicionamiento de un resultado está determinado por el número de enlaces en combinación con el tráfico. Google posiciona mejor las páginas con un elevado número de enlaces y que, al mismo tiempo, tienen un tráfico acorde a los enlaces que recibe. Esta estrategia permite luchar a Google contra la contaminación, es decir páginas que reciben muchos enlaces pero pocas visitas.*

Palabras clave: *PageRank, Visibilidad, Algoritmo de posicionamiento, Alcance, Tráfico.*

Valentín Moreno Pelayo es matemático (especialidad computación), profesor del Departamento de Informática de la Universidad Carlos III de Madrid, investigador en el proyecto «Sistema avanzado de asistencia a la conducción para entornos urbanos: inteligencia artificial" financiado por Cicyt (2004-07), TRA 2004-07441-C03-02/AUT, colaborador en investigación sobre redes temáticas.

Title: Interaction of popularity measurements in web positioning

Abstract: *This work focuses on the relationship between certain popularity measurements and the positioning algorithm of pages retrieved by Google and MSN Search browsers. Measurements studied were: number of page links, number of visitors («reach»), and number of URL requests on a site («pageviews»). The combination of the last two parameters determines page traffic. The following conclusions were drawn from this research: in Google, location of a result, as determined by the number of links, is supported by the page traffic; thus Google performs better in locating those pages in which a high number of links correlates with a high page traffic. This enables Google to avoid contamination: that is, pages with a high number of links and a low number of visits.*

Keywords: *PageRank, Visibility, Positioning algorithm, Reach, Traffic rank.*

Moreno Pelayo, Valentín. *"Interacción entre medidas de popularidad en el posicionamiento web". En: El profesional de la información, 2005, marzo-abril, v. 14, n. 2, pp. 100-107.*

Introducción

La optimización de páginas web para aumentar su visibilidad es un tema clave para los creadores de páginas web[1]. La técnica consiste básicamente en conocer los criterios que utiliza el algoritmo de posicionamiento de cada buscador para mejorar la posición de la página. Uno de los criterios que más se viene utilizando es la medida de la popularidad de determinada página y *Google*[2], con el *PageRank*[3], diseñó un método para medirla basado en el número de enlaces entrantes a la página. Lamentablemente, los webmasters han restado eficacia al aumentar la contaminación mediante la utilización de políticas de intercambios y "granjas de enlaces". Otra forma de medir este dato es calcular el tráfico de visitas que tiene una página (p. e. *Alexa*). En este artículo se estudian posibles relaciones entre:

—Enlaces a la página.

—Cantidad de visitantes.

—Número de peticiones de urls en determinado sitio.

Por lo tanto, se tratará de analizar la relación entre el número de enlaces y el tráfico de determinada página, indicado por los dos últimos factores. Es lógico pensar que a mayor número de enlaces apuntando habrá un mayor número de visitas, por lo que tráfico y enlaces no son variables independientes. En la práctica, las estrategias de optimización hacen que se falsee el número de links, pero la cantidad de visitas es más complicada de manipular. De hecho, un análisis de las páginas indica que la proporción en que se presentan en cada una es en general diferente (a casos con el mismo número de enlaces no les suele corresponder el mismo tráfico y viceversa). Se estudiará si esta rela-

Artículo recibido el 25-09-04
Aceptación definitiva: 11-02-05

ción depende del buscador y cómo afecta al posicionamiento.

En las siguientes secciones se definen, en primer lugar, factores relevantes para el estudio como son el "alcance" y las "páginas visitadas". Posteriormente se realizará un análisis de estos factores basado en los coeficientes de correlación, y por último ofreceremos una posible explicación de los datos observados.

Método

En el presente apartado se describen las siguientes fases:

—Obtención de datos.

—Cálculo de diferentes coeficientes de correlación.

—Comprensión del algoritmo de posicionamiento.

1. Obtención de datos

El análisis se centra en la popularidad. Al no disponer directamente de la información que tienen los buscadores sobre el tráfico, es necesaria una herramienta que aporte datos estadísticos fiables para su estudio. En este caso la aplicación para la obtención de esa información ha sido *Alexa*[4], selección motivada por su fiabilidad y precisión, aunque posee ciertas limitaciones que se comentan a continuación.

Alexa es un sistema de evaluación utilizado y aceptado como parámetro de referencia en el ranking de "popularidad" por los grandes sitios y las grandes empresas en internet. Maneja información sobre sitios web relacionados, estadísticas de visitas, valoración de los usuarios, propietarios, fecha de creación y además realiza comparativas de tráfico con otros sitios analizando semanalmente las tendencias en visitas y páginas vistas.

Posición	Ranking *Alexa*	Puntos por visitas	Enlaces a la página	Puntos enlaces
1	21.905	6	3.384	2
2	22.299	6	767	3
3	27.903	6	214	4
4	308	3	174	4
5	30.334	6	208	4
6	33.669	6	46	5
7	6.060	5	1.066	3
8	33	2	11.706	2
9	12.201	5	39	5
10	99.076	7	1.387	3
11	291	3	1.622	3
12	70.519	6	3	7
13	5.883	5	20	6
14	28.291	6	3	7
15	2.397	4	367	4

Tabla 1. Consulta en Google "adn"

Posición	Ranking *Alexa*	Puntos por visitas	Enlaces a la página	Puntos enlaces
1	6.060	5	1.066	3
2	38.288	6	5	6
3	26.466	6	14	6
4	48.668	6	11	6
5	308	3	174	4
6	291	3	36	5
7	326	3	1	7
8	1.093	4	6.396	2
9	20.164	6	187	4
10	30.334	6	208	4
11	70.519	6	3	7
12	41.689	6	44	5
13	2.251	4	1	7
14	77.399	6	100	5
15	1.311	4	1.831	3
16	5.539	5	294	4
17	2	0	28	5

Tabla 2. Consulta en MSN Search "adn"

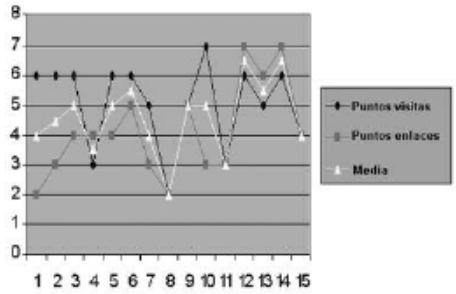


Imagen 1. Gráfica comparativa tráfico/enlaces Google (adn)

a. Ranking de *Alexa*.

a.1. Basado en las visitas de los internautas que tienen instalada su barra (más de 10 millones en todo el mundo) en períodos de tres meses.

Posición	Ranking *Alexa*	Puntos por visitas	Enlaces a la página	Puntos enlaces
1	12.071	5	328	4
2	58.031	6	558	4
3	5.530	5	173	4
4	11.653	5	18	6
5	11.653	5	21	6
6	10.206	5	2.538	3
7	56.663	6	59	5
8	51.110	6	123	5
9	1.696	4	8.938	2
10	38.405	6	61	5
11	2.082	4	428	4
12	11.653	5	24	6
13	3.732	5	689	3
14	16.706	6	124	5
15	1.579	4	10	6
16	52.865	6	320	4
17	52.522	6	614	4
18	3	0	4.423	2
19	5.530	5	173	4
20	48.340	6	7	6
21	23	1	48.562	1
22	1.737	4	2.786	3
23	3.956	5	6.690	2
24	2.957	4	127	4

Tabla 3. Consulta en Google "information retrieval"

a.2. La posición que ocupa un sitio en el ranking mundial es una combinación del alcance y páginas vistas obtenidas, definiéndose estos parámetros como:

—Alcance ("reach"): número de usuarios (direcciones IP) que visitan un sitio en un día dado.

—Páginas visitadas ("page views"): cantidad de páginas visitadas por las urls diferentes que visitan un sitio. En distintos días la misma url se cuenta como diferente.

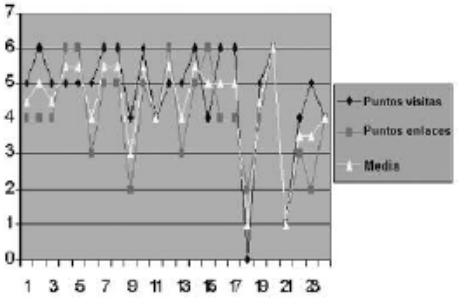


Imagen 3. Gráfica comparativa tráfico-enlaces Google (information retrieval)

b. Sesgos *Alexa*.

b.1. Funcionamiento limitado a los webs de nivel superior (del tipo "www.dominio.com").

b.2. Sólo funciona con el navegador *Internet Explorer* y el sistema operativo *Windows*.

b.3. Los factores culturales y la lengua (la información está en inglés) influyen en la adopción de su software.

b.4. Se desactiva en las páginas seguras (*https:*) de los sitios.

b.5. Los sitios con una posición por encima del puesto 100.000 no son fiables (con menos de 1.000 visitantes mensuales), ya que la cantidad de datos obtenida no es estadísticamente significativa.

2. Desarrollo y resultados

En estudios previos se observó la correlación entre el número de enlaces entrantes y las visitas en la posición de determinada página. Para comprobar este hecho se han realizado búsquedas en *Google* y *MSN Search*. Con un enfoque análogo al propuesto por **Chignell**[5] (1999) se han estudiado los cuarenta y cinco primeros resultados recuperados, frente a los veinte propuestos por este autor, y se han analizado los factores relativos a enlaces entrantes y tráfico.

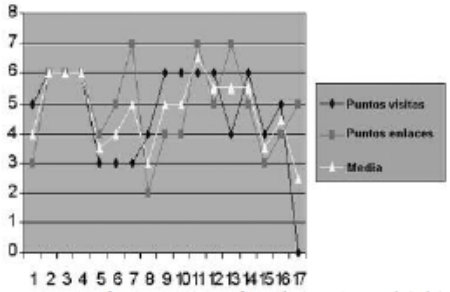


Imagen 2. Gráfica comparativa tráfico-enlaces MSN Search (adn)

Para este artículo se han seleccionado cinco consultas de entre setenta y cinco realizadas con resultados similares: "adn", "information retrieval", "tutorial sql", "BOE vivienda", "tutorial xml". A partir de estas

búsquedas se han analizado más de ochenta páginas en lo relativo a los factores reseñados anteriormente.

2.1. Fases del desarrollo

2.1.1. Obtención de datos

I. Número de resultados por búsqueda y buscador

a. Adn:

—*Google*: 2.600.000.

—*MSN Search*: 1.392.424.

b. Information retrieval:

—*Google*: 6.820.000.

—*MSN Search*: 6.584.613.

c. Tutorial sql:

—*Google*: 2.460.000.

—*MSN Search*: 2.336.830.

d. BOE vivienda:

—*Google*: 125.000.

—*MSN Search*: 25.580.

e. Tutorial xml:

—*Google*: 5.080.000.

—*MSN Search*: 2.755.885.

II. Para cada resultado devuelto en cada búsqueda obtenemos los siguientes datos:

a. Posición entre los resultados de la búsqueda.

b. Número de enlaces a la página.

c. Ranking *Alexa* (posición por tráfico obtenida combinando los criterios siguientes):

—Número de visitantes ("alcance") en tres meses (medidos en millones).

—Cantidad de páginas visitadas (desde ese resultado).

2.1.2. Comprensión del algoritmo de posicionamiento

Para entender cómo afecta la relación entre el tráfico y el número de enlaces de una página en el posicionamiento es conveniente recordar que se puede hacer una aproximación a *Google* para puntuar las páginas en fun-

Posición	Ranking Alexa	Puntos por visitas	Enlaces a la página	Puntos enlaces
1	12.071	5	328	4
2	58.031	6	558	4
3	5.530	5	173	4
4	11.653	5	18	6
5	377	5	10.682	2
6	11.653	5	18	6
7	10.206	5	2.538	3
8	2.902	4	53	5
9	23	1	48.562	1
10	5.530	5	173	4
11	12.071	5	50	5
12	16.706	6	47	5
13	38.405	6	61	5
14	1.864	4	105	5
15	1.239	4	871	3
16	1.696	4	8.938	2
17	1.356	4	144	4
18	8.161	5	22	6
19	4.189	5	3.581	2
20	58.031	6	558	4
21	2.957	4	668	3
22	377	3	10.682	2
23	51.110	6	123	5
24	1.297	4	25	5
25	2.082	4	428	4
26	2.957	4	396	4

Tabla 4. Consulta en MSN Search "information retrieval"

Posición	Ranking *Alexa*	Puntos por visitas	Enlaces a la página	Puntos enlaces
1	1.930	4	1.447	3
2	77.087	6	39	5
3	4.401	5	23	6
4	11.884	5	4.858	2
5	3.386	5	20	6
6	65.854	6	25	5
7	629	4	6	6
8	172	3	16.142	1
9	2.175	4	283	4
10	15.652	6	380	4
11	133	3	318	4
12	66.202	6	21	6
13	52.933	6	88	5
14	25.704	6	3	7

Tabla 5. Consulta en Google "tutorial sql"

Posición	Ranking *Alexa*	Puntos por visitas	Enlaces a la página	Puntos enlaces
1	1.930	4	1.447	3
2	77.087	6	39	5
3	4.401	5	23	6
4	11.884	5	4.858	2
5	94.785	7	21	6
6	65.854	6	25	5
7	133	3	318	4
8	1.489	4	6	6
9	1.930	4	1.447	3
10	4.401	5	23	6
11	77.087	6	39	5
12	3.386	5	20	6
13	2.175	4	283	4
14	126	3	11	6
15	2.306	4	8	6
16	9.086	5	1	7
17	7.849	5	168	4
18	15.036	5	386	4
19	59.036	6	244	4
20	10.757	5	41	5
21	22.506	6	40	5
22	14.195	5	335	4

Tabla 6. Consulta en MSN Search "tuturial sql"

Posición	Ranking *Alexa*	Puntos por visitas	Enlaces a la página	Puntos enlaces
1	47	2	163	4
2	97.488	7	46	5
3	5.087	5	465	3
4	5.381	5	195	4
5	9.718	6	250	4
6	6.783	5	334	3
7	4.828	5	233	4
8	39.086	7	67	4
9	19.699	6	35	5
10	23.070	6	67	4
11	43.864	7	257	4
12	15.989	6	251	4
13	16.372	6	149	4
14	3.339	5	58	4
15	27.932	6	39	5
16	11.179	6	291	3
17	7.856	6	349	3
18	38.851	7	152	4
19	6.021	5	25	5
20	16.372	6	149	4
21	13.624	6	98	4
22	86.375	7	28	5

Tabla 7. Consulta en Google "BOE vivienda"

Puntos	N. de enlaces
1	5
2	25
3	125
4	625
5	3125
6	15625
7	más de 15625

Tabla para la comprensión del algoritmo de posicionamiento

ción del número de enlaces que recibe mediante una escala logarítmica (puede verse un ejemplo en la tabla adjunta dedicada a este punto). Actualmente son al menos diez categorías, y posiblemente se aproxime con un logaritmo distinto del log5, aunque cualitativamente funciona de modo similar.

«Uno de los criterios que más se viene utilizando es la medida de la popularidad de determinada página y *Google*, con el *PageRank*, diseñó un método para medirla basado en el número de enlaces entrantes a la página»

Basándose en esta idea y con el fin de comparar gráficamente los datos de enlaces con los del tráfico, se aplicó el log5 a ambos unificando así la escala en que van a ser representados. Además, sendas gráficas deben de tener el mismo criterio de ordenación respecto al posicionamiento.

Las etapas que se han seguido son:

—Seleccionar sólo las páginas con ranking *Alexa* menor que 100.000 (ver los sesgos de *Alexa*). Cabe destacar que en la mayoría de las búsquedas realizadas las pági-

Posición	Ranking *Alexa*	Puntos por visitas	Enlaces a la página	Puntos enlaces
1	97.488	7	46	5
2	5.087	5	465	3
3	6.783	5	334	3
4	639	4	8	6
5	61.857	7	136	4
6	2.153	5	1.169	3
7	19.516	6	137	4
8	6.783	5	18	5
9	63.206	7	34	5
10	5.087	5	3	6
11	1.929	5	324	3
12	29.035	6	198	4
13	9.718	6	250	4
14	7.856	6	114	4
15	39.086	7	67	4

Tabla 8. Consulta en MSN Search "BOE vivienda"

nas que cumplen este requisito son casi la mitad, una muestra suficientemente significativa ya que partíamos de cuarenta y cinco resultados.

—Obtener la información sobre el número de enlaces y el ranking *Alexa* (tráfico) y aplicarles log5. Los nuevos datos se denominarán puntos enlaces y puntos tráfico respectivamente.

—Cuanto menor sea el valor de puntos tráfico mejor posición, siendo al contrario para puntos enlaces. Reemplazando puntos enlaces por siete menos puntos enlaces invertimos el criterio de ordenación facilitando así el estudio de la relación entre ambos criterios.

—Construir gráficas comparativas para puntos enlaces y puntos tráfico por búsqueda y buscador.

Conclusiones

Aunque es lógico pensar que a mayor número de enlaces más visitas, el estudio muestra evidencia de que la correlación entre uno y otro no es perfecta. También se han encontrado diferencias entre *Google* y *MSN Search*. En el primero el posicionamiento de un resultado determinado por el número de enlaces es avalado por el tráfico en el mismo. El coeficiente de correlación está próximo a 0,5 de media, suficientemente significativo si consideramos la gran cantidad de factores que influyen en el posicionamiento.

Se observa que las gráficas realizadas con los datos de los puntos por enlace y los puntos por tráfico son cualitativamente similares, es decir, *Google* posiciona mejor la página cuyo número de enlaces a la misma es elevado y además acorde al tráfico que recibe. Esto le permite luchar contra la contaminación (las páginas que reciben muchos enlaces pero pocas visitas no se van a posicionar bien). Es sencillo incluir esta idea en una fórmula: es suficiente con obtener la posición de una página considerando el criterio del número de enlaces, conseguir su posición atendiendo ahora al criterio de tráfico recibido y por último calcular la distancia entre ambas. Si esta diferencia es pequeña su posicionamiento se verá favorecido.

MSN Search también utiliza estos dos criterios para posicionar pero sin exigir concordancia entre ellos

Posición	Ranking *Alexa*	Puntos por visitas	Enlaces a la página	Puntos enlaces
1	1.943	5	1.447	2
2	721	4	5.946	2
3	78.874	7	760	3
4	21.084	6	1.333	3
5	56.116	7	440	3
6	10	1	4.868	2
7	721	4	5.946	2
8	9.596	6	1.326	3
9	39.494	7	946	3
10	50.719	7	132	4
11	39.494	7	946	3
12	1.473	5	687	3
13	10.770	6	468	3
14	333	4	7.822	1
15	64.074	7	498	3
16	333	4	7.822	1
17	64.074	7	498	3
18	89.993	7	386	3
19	2.356	5	1.186	3
20	1.479	5	10.686	1
21	8.886	6	2.360	2
22	8.638	6	584	3
23	22.349	6	1.507	2

Tabla 9. Consulta en Google "tutorial xml"

Posición	Ranking *Alexa*	Puntos por visitas	Enlaces a la página	Puntos enlaces
1	1.943	5	1.447	2
2	721	4	5.946	2
3	10	1	4.868	2
4	56.116	7	440	3
5	21.084	6	1.333	3
6	1.479	5	10.639	1
7	21.680	6	921	3
8	89.993	7	386	3
9	8.638	6	584	3
10	1.463	5	6	6
11	65.611	7	965	3
12	9.887	6	1.109	3
13	18.454	6	953	3
14	25.926	6	98	4
15	63.004	7	104	4
16	1.070	4	432	3
17	89.993	7	386	3

Tabla 10. Consulta en MSN Search "tutorial xml"

(en varias búsquedas el coeficiente de correlación está muy próximo a cero). En las gráficas se observa que posiciona bien las páginas con muchos enlaces y en proporción menos tráfico, y viceversa. Si una página destaca en uno de los criterios puede compensar un resultado modesto en el otro.

«En la práctica, las estrategias de optimización hacen que se falsee el número de links, pero la cantidad de visitas es más complicada de manipular»

Así, se puede concluir que aunque *Google* da más importancia a los enlaces entrantes, algo corroborable por la

	Adn	Information retrieval	Tutorial sql	BOE vivienda	Tutorial xml
Google	0,307	0,589	0,539	0,30	0,768
MSN Search	0,081	0,521	0,135	–0,013	0,345

Tabla 11. Coeficientes de correlación (tráfico/enlaces)

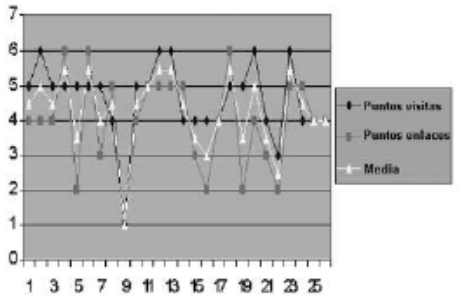


Imagen 4. Gráfica comparativa tráfico-enlaces MSN Search (information retrieval)

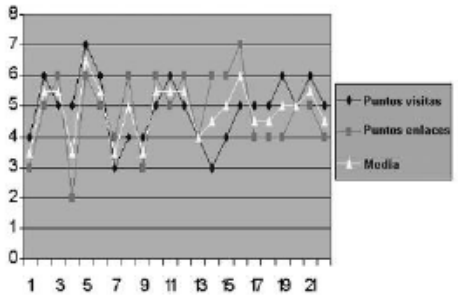


Imagen 6. Gráfica comparativa tráfico-enlaces MSN Search (tutorial sql)

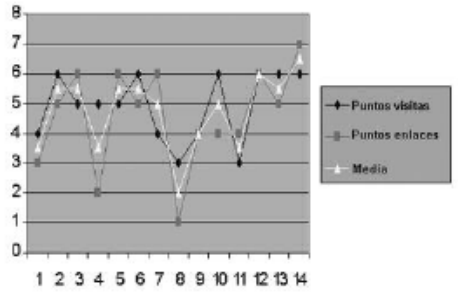


Imagen 5. Gráfica comparativa tráfico-enlaces Google (tutorial sql)

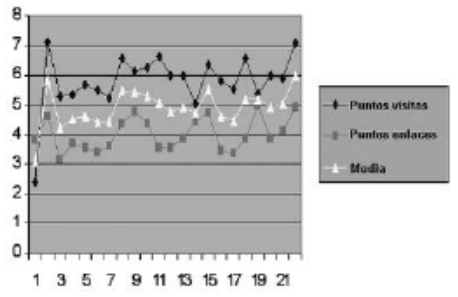


Imagen 7. Gráfica comparativa tráfico-enlaces Google (BOE vivienda)

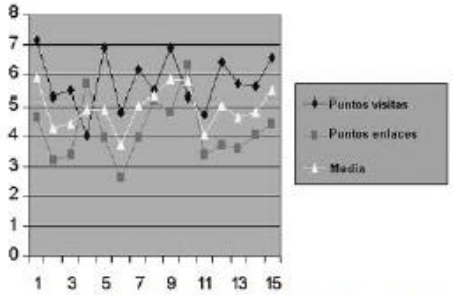


Imagen 8. Gráfica comparativa tráfico-enlaces MSN Search (BOE vivienda)

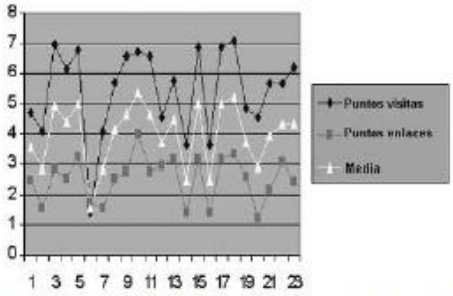


Imagen 9. Gráfica comparativa tráfico-enlaces Google (Tutorial xml)

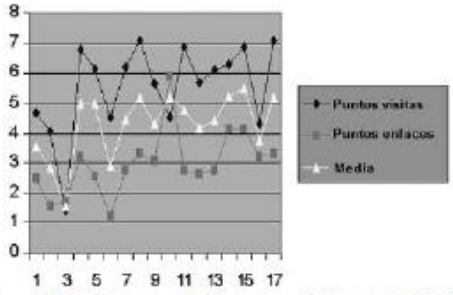


Imagen 10. Gráfica comparativa tráfico-enlaces MSN Search (tutorial xml)

estabilidad de sus resultados, debe estar teniendo en cuenta de alguna forma el número de visitas. Esta circunstancia no se da en *MSN Search*, lo que creemos que es algo consciente y no una consecuencia indirecta de la dependencia entre tráfico y enlaces ni debido a que contabiliza de igual forma los vínculos entrantes de páginas mediocres y los de elevada calidad.

Otras observaciones relevantes extraídas son:

—La relación entre el alcance y las páginas visitadas y tráfico es muy significativa: coeficientes de correlación próximos a 1, en torno a 0,998.

—*Google* recupera un número significativamente mayor de documentos (en algunas búsquedas casi el doble).

—En el cálculo del tráfico se da casi el doble de peso al alcance que a las páginas visitadas (según lo realiza *Alexa*).

—*MSN Search* y *Google* recuperan en las primeras posiciones un número similar de documentos por debajo de la posición 100.000 de la clasificación por tráfico que hace *Alexa*.

Bibliografía

1. **Diligenti, Michelangelo; Fellow, Marco Gori; Maggini, Marco.** "A unified probabilistic framework for web page scoring systems". En: *IEEE transactions on knowledge and data engineering*, 2004, January, v. 16, n. 1, pp. 4-16.

2. *Google. Google PageRank, GoogleMania, Todo sobre Google*. Consultado en: 01-10-04.
http://www.googlemania.com/pagerank.php

3. **Page, L.; Brin, S.; Motwani, R.; Winograd, T.** "The PageRank citation ranking: bringing order to the web". Technical report, Computer Science Dept., Stanford Univ., 1998.

4. *Alexa Internet, Inc. Alexa Web Search. About the Alexa traffic rankings*. Consultado en: 01-10-04.
http://pages.alexa.com/prod_serv/traffic_learn_more.html

5. **Chignell, M. H.; Gwizdka, J.; Bodner, R. C.** "Discriminating meta-search: a framework for evaluation". En: *Information processing and management*, 1999, v. 35, n. 3, pp. 337-362.

***Valentín Moreno Pelayo**, Departamento de Informática, Universidad Carlos III de Madrid.*
vmpelayo@inf.uc3m.es

Anexo B: Herramienta desarrollada ad hoc

Esta sección recoge aspectos relacionados con la implementación de la aplicación desarrollada para la experimentación, como las decisiones que se han tomado sobre herramientas de programación, lenguaje de programación o diseño de soluciones. Por otra parte, también se discuten las características especiales del programa con el fin de mostrar brevemente cómo se han resuelto algunos de los problemas encontrados.

B.1 Metodología de la Programación

Para la aplicación desarrollada con el fin de realizar y evaluar la experimentación, se ha elegido la orientación a objetos como paradigma de programación. En concreto, la herramienta es una aplicación de escritorio en lenguaje Java. Para el desarrollo de la aplicación se escogió NetBeans como entorno de desarrollo. Las principales ventajas que han sido consideradas para su elección se resumen en:

- Posibilidad de crear una interfaz amigable para el usuario por medio de ventanas.
- Capacidad de ejecutar el software en cualquier plataforma, aumentando por tanto el número de usuarios potenciales.
- Libre distribución tanto de Java como de los entornos de desarrollo Eclipse o NetBeans (Knudsen y Niemeyer, 2005).

B.2 Funcionalidades Generales

El programa está obligado a proporcionar las siguientes funcionalidades generales:

- **Independencia de los procesos de la aplicación: C**ada proceso puede tardar mucho tiempo en su ejecución. Por esta razón, la mejor opción es permitir al usuario ejecutar cada uno de ellos por separado.

- **Configuración de la ejecución:** Para las funcionalidades que requieren más recursos, se podrá limitar algunos aspectos de los parámetros de ejecución de forma que la aplicación puede funcionar más rápido. Por ejemplo, estableciendo el número máximo de enlaces a analizar o el tamaño máximo de una página web para realizar su lectura. El usuario de esta forma puede alcanzar un compromiso entre el rendimiento y la precisión de los resultados.
- **Tratamiento de frases exactas:** La herramienta debe permitir realizar análisis de consultas de frases exactas. Es decir, los resultados de estas consultas contienen los términos de consulta consecutivos y en el mismo orden.
- **Consultas en varios idiomas:** El programa apoyará el análisis de los resultados de búsqueda relacionados con las consultas en inglés y español.

Las funcionalidades concretas se describen siguiendo los módulos en que se ha dividido la aplicación. Los módulos son:

1. **Analizador de resultados de búsqueda**: El objetivo de ese módulo es capturar los datos de los resultados asociados a consultas realizadas en los motores de búsqueda.
2. **Generación de modelos:** Este módulo es el encargado de crear los modelos de estimación del posicionamiento a partir de los datos recogidos por el módulo anterior.
3. **Estimador:** En este módulo se llevan a cabo los pronósticos de estimación de relevancia de páginas web, se aplican para ello los modelos generados en el módulo anterior de generación de modelos.

B.3 Módulo Analizador de resultados de búsqueda

El módulo *Analizador de resultados de búsqueda* permite realizar consultas en los motores de búsqueda con los que se ha experimentado. Las páginas web obtenidas como resultados son procesadas con el fin de obtener los valores correspondientes de las variables SEO analizadas.

A continuación para mostrar el modelado del Analizador de resultados de búsqueda se muestra un diagrama de clases UML (Rumbaugh et al., 1998). En este diagrama de clases se describen las principales relaciones entre las clases del módulo *Analizador de*

resultados de búsqueda. Las clases de interfaz de usuario no han sido incluidas para simplificar el diseño y su comprensión.

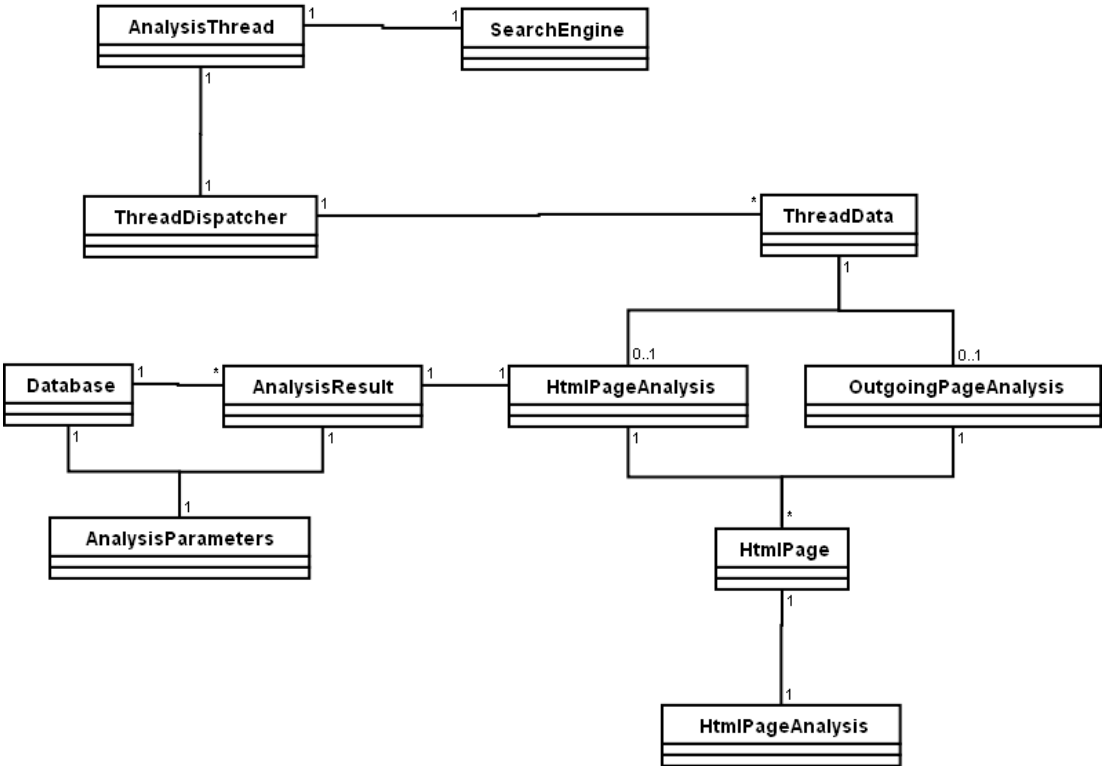


Figura B-1 Diagrama de clases del módulo Analizador de resultados de búsqueda

Se describen a continuación cada una de las clases y sus relaciones:

SearchEngine: Contiene todos los métodos necesarios para ejecutar consultas en los motores de búsqueda Google, Yahoo Search y MSN. También proporciona un método para obtener la dirección URL de todos los enlaces entrantes de una página web determinada, utilizando la base de datos de Google. Todos los métodos (subprogramas) proporcionados por esta clase son estáticos.

Relaciones: Esta clase se utiliza en dos casos específicos: cuando se obtiene la lista de sitios web que tienen que ser analizados según las palabras clave introducidas en la aplicación, y en la obtención de los enlaces entrantes de una página web.

AnalysisThread: Es el hilo principal de la aplicación, su principal objetivo es la obtención de la lista de las páginas web a analizar, para pasarlos a la clase *ThreadDispatcher* y esperar hasta que esta última haya terminado de analizar

todas ellas. Fue diseñada como un hilo para permitir una integración más fácil en la interfaz de usuario.

Relaciones: Esta clase, utiliza la clase *SearchEngine* para obtener los resultados asociados a las palabras clave, y la clase *ThreadDispatcher* para iniciar y controlar el análisis de los sitios web resultantes.

ThreadDispatcher: Crea y gestiona los hilos que analizan las páginas web. El objetivo principal de esta clase es hacer un seguimiento de los hilos en ejecución y de los que están en espera, controlando los hilos que pueden actuar de forma simultánea. De esta manera, el número de subprocesos no crece de manera exponencial y los recursos informáticos no quedan sobrecargados.

Relaciones: La clase *ThreadDispatcher* es utilizada por el hilo principal de la aplicación, como se ha explicado antes. Tiene dos listas de datos en espera para ser analizados: una para el análisis de las páginas web, y otra para el análisis de los enlaces entrantes y salientes.

ThreadData: Esta clase encapsula información sobre las páginas web que tienen que ser analizadas. Contiene todos los parámetros necesarios para permitir que la clase *ThreadDispatcher* cree el hilo correspondiente e inicie el análisis de la página, es decir: URL, palabras clave, tipo de hilo, posición que ocupa entre los resultados de búsqueda y un valor booleano que indica si las palabras clave deben ser manejadas como una frase exacta.

Relaciones: Esta clase se almacenan en dos listas por la clase *ThreadDispatcher*, tal como se explicó. Como hay dos tipos diferentes de análisis (para páginas web y los enlaces entrantes o salientes), la clase *ThreadData* puede estar relacionada con el análisis de la página principal o con el análisis de páginas externas.

HtmlPage: Esta clase encapsula métodos de lectura y análisis del contenido de páginas HTML. Además, la clase proporciona métodos para obtener información del documento HTML, como por ejemplo: el texto sin formato o el contenido HTML en una etiqueta determinada, los parámetros en una declaración de etiqueta, el valor de un atributo dado de una etiqueta, etc.

Relaciones: Los objetos *HtmlPage* se utilizan durante el proceso de análisis para extraer información de la página web.

HtmlAnalysis: Esta clase ofrece numerosos métodos estáticos para el cálculo de los diferentes parámetros de análisis de una página HTML. Todos ellos reciben como entrada la página HTML analizada, así como otros parámetros importantes (por ejemplo: palabras clave, la posición de resultado de búsqueda, etc.).

Relaciones: Los métodos de análisis se realizan en objetos de la clase *HtmlPage.*

MainPageAnalysis: Esta clase contiene los resultados de los análisis realizados sobre las páginas web obtenidas como resultados de consultas sobre un motor de búsqueda. Está diseñado como un hilo, que es creado y controlado por la clase *ThreadDispatcher*. Uno por uno, el método principal de este hilo ejecuta el análisis de todos los parámetros, almacena sus valores y crea un objeto de la clase *AnalysisResult*. También ofrece cálculos especiales de parámetros para las páginas con frames.

Relaciones: El hilo principal de la clase *MainPageAnalysis* es administrado por la clase *ThreadDispatcher*, y por lo tanto depende de ella. Para realizar el análisis, primero es necesario leer la página HTML que tiene que ser procesada. A continuación, los métodos de análisis de cada parámetro se ejecutan y los resultados se almacenan en un objeto de la clase *AnalysisResult*.

ExternalPageAnalysis: Esta clase contiene los resultados de los análisis realizados sobre páginas web vinculadas a una página principal. Como en el caso de la clase anterior, se ha diseñado como un hilo, que es creado y controlado por la clase *ThreadDispatcher*. La principal diferencia entre este análisis y el realizado en las páginas principales es que éste sólo busca en el contenido de la página las palabras clave y no analiza los demás parámetros. Es decir, el propósito final es saber si una página contiene las palabras clave o no, ya sea en el título, en la dirección o en el cuerpo del texto.

Relaciones: El hilo de *ExternalPageAnalysis* es administrado por el *ThreadDispatcher*. Los métodos de la clase analizan tres parámetros para determinar si las palabras clave aparecen en el título, en la dirección o en el texto del cuerpo de la página web.

AnalysisResult: Esta clase contiene todos los valores de los parámetros que se han obtenido al analizar una página web principal. El propósito de encapsularlos es simplificar el almacenamiento de esos resultados en la base de datos.

Relaciones: Los resultados del análisis se obtienen por la clase *MainPageAnalysis*, y posteriormente son almacenados por un método de la clase *Database*.

Database: Esta clase contiene todos los métodos estáticos que son necesarios para cargar y almacenar la información en una base de datos Access. Además, también ofrece un método para eliminar los resultados de un análisis específico.

Relaciones: Los resultados de *Database* obtenidos de los parámetros del programa se pueden pasar a la clase *AnalysisParameters*. También almacena los resultados de un análisis, tomando los valores almacenados dentro de un objeto de la clase *AnalysisResult*.

AnalysisParameters: Esta clase contiene todos los parámetros configurables por el usuario en los análisis, tales como tiempos de espera, o la decisión de analizar los enlaces entrantes y salientes. Estos valores son accesibles desde casi cualquier clase del programa, ya que se necesitan para determinar el modo en qué tiene que ser ejecutado.

Relaciones: Los parámetros se cargan desde la base de datos y posteriormente se utilizan por las clases del programa.

B.4 Módulo de Generación de Modelos

Esta sección describe los aspectos más importantes relacionados con el módulo de *Generación de modelos*. Este módulo se diferencia de los demás porque delega tareas de importancia a un programa externo, Weka. El propósito de este módulo es preparar los datos de entrada para Weka y ejecutar esta herramienta desde la aplicación.

B.4.1 Funcionalidades

El módulo *Generador de modelos* se ha diseñado para que pueda proporcionar las siguientes funcionalidades:

- **El uso de varios algoritmos de clasificación:** El módulo es capaz de utilizar varios algoritmos de clasificación. Los algoritmos de clasificación son comparados con el

fin de especificar cuál de ellos es el más adecuado para la resolución del problema de estimación del posicionamiento.

- **Posibilidad de exportar los modelos:** Se podrá tomar cada modelo individual generado por Weka y exportarlo, de modo que pueda ser utilizado por la aplicación Weka para clasificar cualquier otra instancia de páginas web.

B.4.2 Conexión con Weka

Weka está programado en Java y se distribuye como una aplicación de Windows. Sin embargo, el núcleo del programa es un archivo JAR llamado "weka.jar", que puede obtenerse de la distribución GNU de código abierto (Stallman et al., 2004). Este archivo contiene todas las clases binarias necesarios para ejecutar Weka y por tanto, cualquier algoritmo de análisis incluido en esta herramienta. No es necesario hacer llamadas externas a Weka, y es suficiente con vincular el archivo JAR y utilizar su API.

Mediante la API de Weka se han llamado a los algoritmos utilizados en la experimentación. En la siguiente tabla se muestran las líneas de comandos correspondientes con cada algoritmo de clasificación.

Algoritmos	Líneas de comandos
C4.5	`-o -C 0.25 -M 2 -t salida.arff -d out.model`
PART	`-o -C 0.25 -M 2 -Q 1 -t salida.arff -d out.model`
Bagging C4.5	`-o -P 100 -S 1 -I 10 -W weka.classifiers.trees.J48 -t salida.arff -d out.model -- -C 0.25 -M 2`
Bagging PART	`-o -P 100 -S 1 -I 30 -W weka.classifiers.rules.PART -t salida.arff -d out.model -- -C 0.25 -M 2 -Q 1`
Boosting C4.5	`-o -P 100 -S 1 -I 10 -W weka.classifiers.rules.PART -t salida.arff -d out.model -- -C 0.25 -M 2`
Boosting PART	`-o -P 100 -S 1 -I 10 -W weka.classifiers.rules.PART -t salida.arff -d out.model -- -C 0.25 -M 2 -Q 1`

Tabla B-1: Líneas de comandos correspondientes a los algoritmos de clasificación

B.4.3 Archivos de entrada de datos

Para construir un modelo de decisión, Weka tiene que tomar como entrada un conjunto de instancias. Estas instancias deben ser almacenadas en un archivo ARFF (Hall et al.,

2009). Los archivos ARFF son archivos de texto ASCII que describen una lista de ejemplos de aprendizaje que comparten un conjunto de atributos. El formato de estos archivos es fijo, y tiene que respetarse escrupulosamente para que Weka pueda interpretar los datos contenidos en el mismo.

Los archivos ARFF tienen dos partes bien diferenciadas. La primera sección contiene la información de encabezado, que incluye el nombre de la relación, una lista de los atributos (las columnas en las instancias de datos), y sus tipos. La siguiente ilustración muestra un ejemplo de encabezado de este tipo de archivos.

```
@relation betterpositioned

@attribute percentage_kw_title_A real
@attribute percentage_kw_body_A real
@attribute number_links_incoming_A real
@attribute percentage_kw_title_B real
@attribute percentage_kw_body_B real
@attribute number_links_incoming_B real
@attribute class {0, 1}
```

Figura B-2: Ejemplo de encabezado de archivos ARFF

La segunda sección contiene los datos en sí. Cada instancia está representada en una línea independiente con los atributos separados por comas. El número y posición de cada valor del atributo debe coincidir con la descripción que figura en el encabezado del archivo ARFF. Un ejemplo de esta sección de datos se muestra en la siguiente ilustración.

```
@data
0.0,0.0,0.0,50.0,14.0,0.0,0
100.0,0.0,0.0,82.75,85.0,0.0,1
100.0,0.0,0.0,81.5,70.0,0.0,0
100.0,0.0,0.0,82.75,85.0,0.0,1
```

Figura B-3: Ejemplo de datos de archivos ARFF

Como se explicó en el marco experimental, los ejemplos de aprendizaje que reciben los algoritmos primero tendrán los valores de ciertas variables de una página web, seguidos de los valores de las mismas variables de otra página web y, por último, el atributo de clase, que indica cual de las dos páginas web está mejor posicionada (ver Figura III-6).

B.4.4 Modelo generado

Después de ejecutar un análisis en Weka, se crea un archivo con el modelo resultante. Este modelo puede ser utilizado para clasificar nuevas instancias. Por lo tanto, es

necesario que la herramienta desarrollada almacene estos archivos para posteriores fases de estimación del posicionamiento.

Los modelos son almacenados en una carpeta llamada "wekamodels", por lo que pueden ser encontrados por el módulo *Estimador*. Es necesario adjuntar información adicional que permita identificar los modelos y seleccionar los más convenientes para hacer pronósticos de posicionamiento. Debido a esta necesidad, la siguiente información se incluye en el nombre de los archivos:

- Algoritmo utilizado para generar el árbol de decisión (es decir, C4.5, PART, etc.).
- Motor de búsqueda de los que provienen los datos con los que se ha construido el modelo.
- Tasa de éxito del modelo.
- Variables utilizadas en la generación del modelo.
- Variables que fueron normalizadas.

Esta información está codificada de forma abreviada en el nombre de archivo con el fin de no exceder el número máximo de caracteres permitidos para el nombre. Otra ventaja de este sistema de nomenclatura es que permite disponer de varios modelos de clasificación diferentes sin tener que sobrescribir los ya existentes.

B.4.5 Redirección de la salida estándar

Además de generar un archivo de modelo de salida, Weka muestra información sobre el proceso de análisis a través de la salida estándar del sistema. Los datos que se exhiben por pantalla contienen información importante para los propósitos de este trabajo, tales como la tasa de éxito del modelo creado, o la matriz de confusión. La siguiente ilustración se corresponde con una salida de datos por pantalla de la herramienta Weka.

```
J48 pruned tree
------------------

number_links_incoming_A <= 14
|  percentage_kw_body_A <= 6.100796
|  |  number_links_incoming_B <= 14
|  |  |  percentage_kw_body_B <= 6.100796
……
|  number_links_classified_0_A > 14: 0 (198.0/40.0)

Number of Leaves :  411

Size of the tree :  821


Time taken to build model: 3.42 seconds
Time taken to test model on training data: 0.05 seconds

=== Error on training data ===

Correctly Classified Instances        9184               92.7677 %
Incorrectly Classified Instances       716                7.2323 %
Kappa statistic                          0.8554
Mean absolute error                      0.1192
Root mean squared error                  0.2441
Relative absolute error                 23.8321 %
Root relative squared error             48.8182 %
Total Number of Instances             9900


=== Confusion Matrix ===

   a   b   <-- classified as
 4615 335 |   a = 0
 381 4569 |   b = 1



=== Stratified cross-validation ===

Correctly Classified Instances        8547               86.3333 %
Incorrectly Classified Instances      1353               13.6667 %
Kappa statistic                          0.7267
Mean absolute error                      0.1703
Root mean squared error                  0.3347
Relative absolute error                 34.0569 %
Root relative squared error             66.9321 %
Total Number of Instances             9900


=== Confusion Matrix ===

   a   b   <-- classified as
 4261 689 |   a = 0
 664 4286 |   b = 1
```

Figura B-4: Weka Ejemplo de salida

Estos datos también son capturados por la aplicación que se ha desarrollado mediante una redirección de la salida estándar de Weka.

B.5 Módulo Estimador

El módulo *Estimador* está obligado a proporcionar las siguientes funcionalidades:

- **Posibilidad de elegir entre varios modelo de estimación:** El usuario podrá elegir entre todas los modelos que se ha generado previamente el módulo anterior. El sistema deberá indicar con claridad la fiabilidad de cada uno de ellos, por lo que el usuario podrá elegir el que mejor se adapte a sus necesidades.
- **Comparar con las páginas web más competitivas:** El módulo *Estimador* deberá ser capaz de dar consejos al usuario sobre cómo mejorar la posición de su sitio web. Informará al usuario sobre los valores de las variables SEO más significativas de su sitio web y los comparará con los de las páginas web mejor posicionadas.
- **Proporcionar medidas de fiabilidad:** La fiabilidad de la posición estimada depende del modelo de estimación. Es necesario informar al usuario sobre la fiabilidad de la estimación, proporcionando variables estadísticas del éxito del modelo en las predicciones.
- **Estimación del posicionamiento de páginas web locales:** La herramienta debe permitir el análisis de páginas web locales. Esta funcionalidad está concebida para comprobar si páginas que aún no están en Internet tiene un contenido adecuado.

Diagrama de clases

En la Figura B-5 se presenta el diagrama de clases UML (Rumbaugh et al., 1998) del módulo *Estimador*. Se detallan las clases y las relaciones entre ellas con el fin de establecer el diseño en el que se basa la implementación.

Este módulo está muy relacionado con el módulo *Analizador de resultados de búsqueda*, ya que una de las principales tareas del estimador es analizar el contenido de la página web cuya posición quiere ser estimada. Por lo tanto, no se vuelven a detallar las clases coincidentes en ambos diagramas.

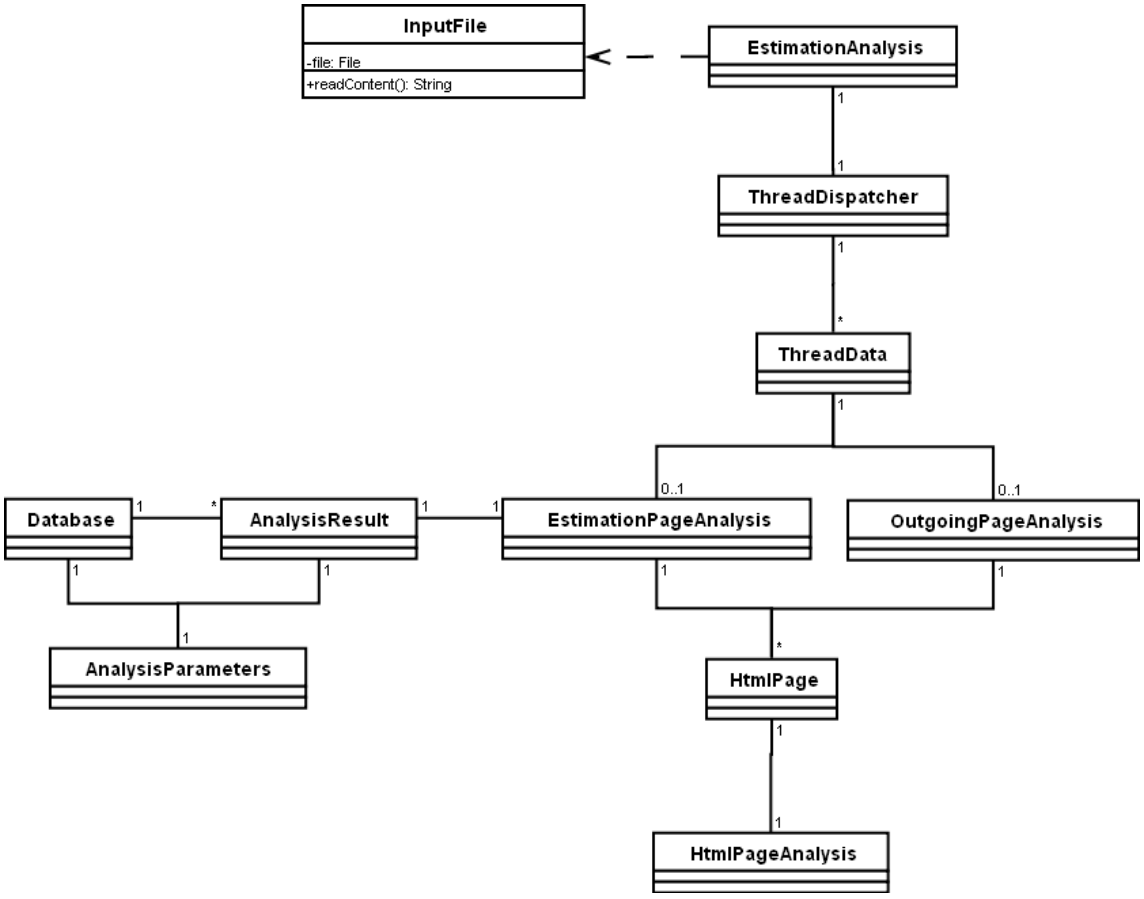


Figura B-5: Diagrama de clases del módulo Estimador

La descripción de las clases exclusivas de este módulo y las relaciones entre ellas se presenta a continuación:

- **InputFile:** Esta clase se utiliza para leer el contenido de archivos HTML locales. La ruta de acceso local del archivo HTML, que se desea utilizar, permite leer su contenido y retornarlo en forma de cadena de caracteres.

  Relaciones: Esta clase es utilizada por la clase principal del módulo *Estimador*, *EstimationAnalysis*. Este último se encarga de detectar cuando el usuario ha seleccionado un archivo local para ser analizado, leyendo su contenido con la ayuda de esta clase.

- **EstimationAnalysis:** Esta clase se encarga de organizar el análisis de la página web cuya posición quiere ser estimada. Proporciona una estimación de la posición que ocuparía la página web entre los resultados de una consulta.

  Relaciones: Como se explicó anteriormente, *EstimationAnalysis* utiliza la clase *InputFile* para acceder al contenido de archivos locales. En cualquier caso, es también responsable

de la inicialización y el control de la clase *ThreadDispatcher*, a fin de analizar la página web de entrada.

- **EstimationPageAnalysis:** Esta clase contiene los resultados del análisis realizado en la página web cuya posición se está estimando. Está diseñado como un hilo, que es creado y controlado por el *ThreadDispatcher*. El método principal de este hilo ejecuta el análisis de todos los parámetros almacenando su valor con el fin de crear un objeto *AnalysisResult*. También ofrece cálculos de parámetros especiales para las páginas con frames (ver apartado 3.2.2.1, II, C).

  Relaciones: Esta clase depende de la clase *ThreadDispatcher*. Además los métodos de análisis de cada parámetro almacenan los resultados en un objeto *AnalysisResult*.

B.6 Interfaz de usuario

La interfaz de usuario proporciona las siguientes funcionalidades:

- **Interfaz gráfica de usuario:** Es una interfaz amigable distinguiendo claramente los tres principales funcionalidades del programa (análisis de resultados de búsqueda, generación de modelos de estimación y la estimación de la posición).
- **Información sobre el estado actual análisis:** Se muestra mediante una barra de progreso para informar al usuario sobre el estado de la ejecución.
- **Integración de todas las herramientas en el marco del programa:** Todos las herramientas utilizadas están integradas en el entorno de la aplicación. Esto simplifica el uso del programa y le aporta mayor coherencia.
- **Información de los errores del proceso:** Los errores se muestran claramente y son explicados a través de la interfaz de usuario.

  En el entorno NetBeans, la GUI (*Graphical User Interface*) por defecto es Swing (Elliott y Eckstein, 2002), que es parte de JFC (*Java Foundation Classes*). El conjunto de herramientas Swing incluye un conjunto de componentes para la construcción de interfaces gráficas de usuario y añade interactividad a las aplicaciones Java. Swing incluye todos los componentes que normalmente se añaden a la interfaz de ventanas.

  Las interfaces de la aplicación son simples gráficos 2D. La siguiente ilustración (Figura B-6) presenta la mayoría de los componentes que son utilizados en cualquier interfaz del programa.

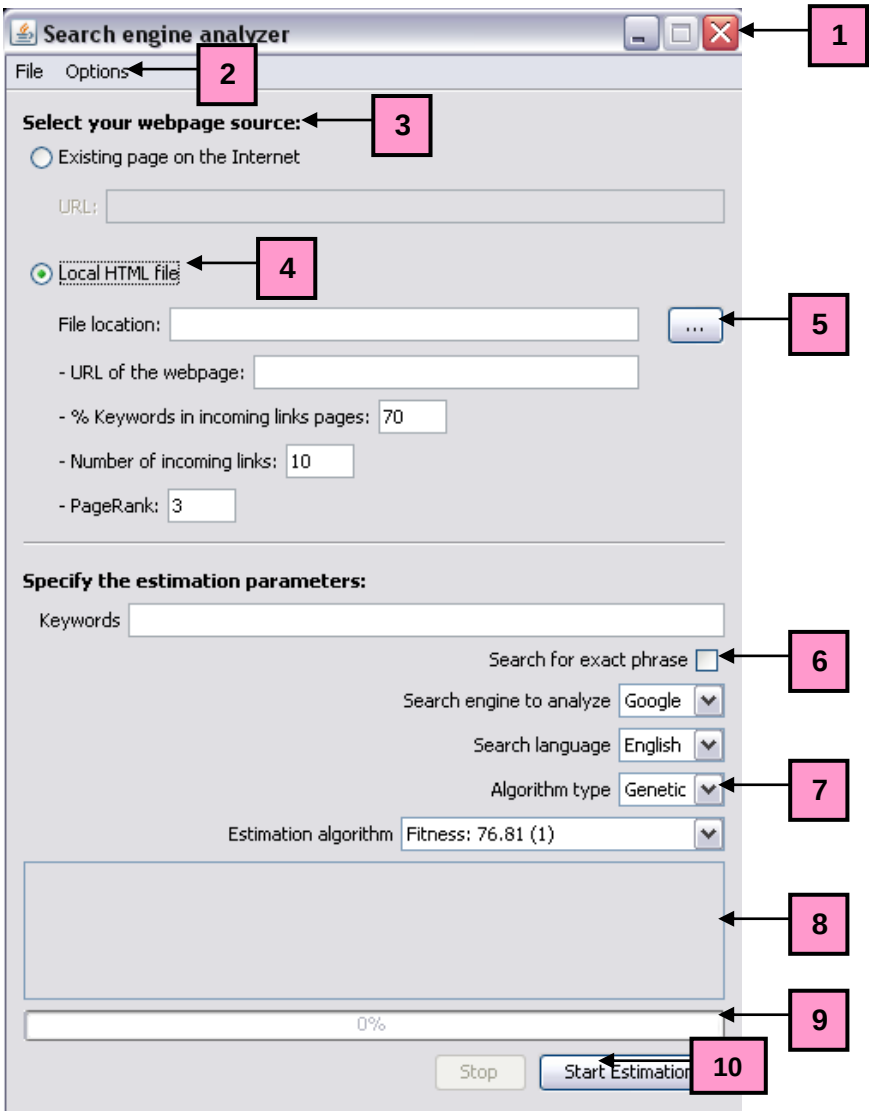


Figura B-6: Principales componentes de la interfaz

Se explica a continuación cada una de las componentes. Destacar que no hay ninguna diferencia de uso de un mismo componente entre las diversas interfaces del programa, ya que cada uno de los componentes se utiliza para los mismos fines en cualquiera de ellas.

1. **Botones por defecto:** En todas las ventanas se muestran los botones por defecto que sirven para minimizar, maximizar y cerrar. La operación de cierre no está habilitada en las ventanas en las que se requiere una confirmación por parte del usuario.

2. **Barra de menú:** La barra de menú se utiliza para proporcionar la interfaz de opciones avanzadas mediante una nueva ventana. La opción de salida se incluye, también, cada vez que la barra de menú se muestra.

3. **Etiquetas de texto:** Aparecen constantemente en la ventana de interfaces de la aplicación ya que se utilizan para mostrar textos informativos para el usuario. En estas etiquetas de texto sólo se mostrará información estática. Toda la información generada por la aplicación se mostrará en las áreas de texto o en cuadros de diálogo de mensajes.
4. **Botones de radio:** Se muestran en los casos en que existen varias opciones mutuamente excluyentes.
5. **Diálogos de selección de archivos:** Este tipo especial de cuadro de diálogo se utiliza para mostrar una ventana de exploración de archivos, donde el usuario puede navegar a través de su sistema de archivos con el propósito de seleccionar uno.

Figura B-7: Diálogo de selección de archivos

6. **Casillas de verificación:** Este control se utiliza para seleccionar opciones de aplicación. Si una casilla está marcada significa que el usuario desea activar esa opción, en caso contrario no se aplicará.
7. **Los cuadros combinados:** Este tipo de componentes permite al usuario seleccionar una de varias opciones mediante desplegables.
8. **Área de texto:** Esta parte de la interfaz muestra información de la aplicación tras su ejecución. Su objetivo principal es mostrar mensajes informativos, ya sea acerca de los errores, advertencias o mensajes de éxito.
9. **Barra de progreso:** Esta barra muestra el progreso de la ejecución del programa.

10. **Botones:** Estos componentes se utilizan principalmente para permitir al usuario iniciar y detener un proceso de análisis. Si la tarea se está ejecutando el botón de inicio estará deshabilitado y el botón de parada activado, y viceversa.

Adicionalmente a los componentes de interfaz típicos y las ventanas se utilizan otros componentes especiales, tales como mensajes de error y cuadros de diálogo:

11. **Mensajes de dialogo:** Estos mensajes se utilizan para mostrar información o advertencias sobre la ejecución del programa. Su uso más común es mostrar el resultado final de una tarea de análisis. En algunos casos no ofrecen ninguna elección posible para el usuario, pero en otros casos se les pide una respuesta, como en la ventana de la siguiente ilustración:

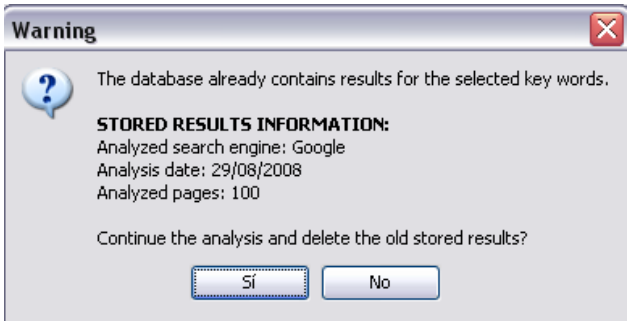


Figura B-8: Mensaje de diálogo

12. **Mensajes de error:** Se utilizan en los casos en que se produce un error importante. Por ejemplo, el mensaje de error siguiente se muestra cuando la base de datos es inaccesible para el programa.

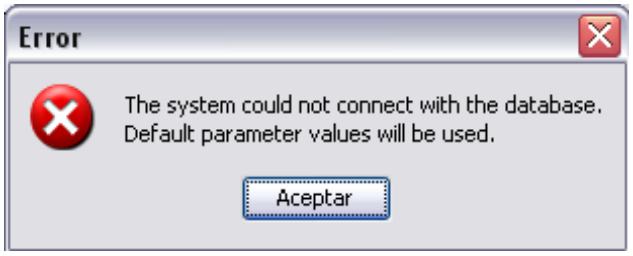


Figura B-9: Mensaje de error

B.7 Manual del Usuario

Esta sección tiene como objetivo explicar cómo utilizar la aplicación desarrollada con el fin de poder obtener la posición estimada de una página web. Por lo tanto, se explica el conjunto de pasos que se tienen que seguir para su explotación.

Como ejemplo, vamos a tratar de conocer la posición de una página web, que queremos optimizar con las palabras clave "computer engineering" y "genetic algorithms", para el motor de búsqueda de Google.

Paso 1: Resultados de las consultas a analizar

El primer paso consiste en realizar una consulta de búsqueda de algunas palabras clave en el motor de búsqueda deseado. Seleccionamos la primera opción del menú principal:

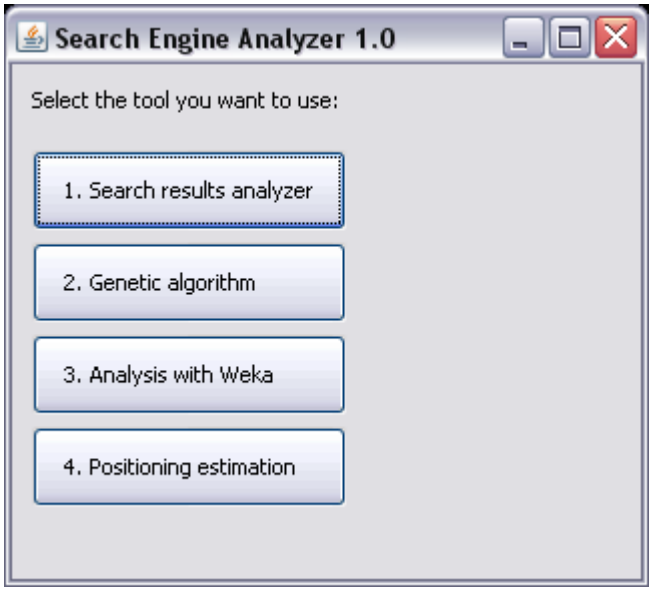


Figura B-10: Interfaz del menú principal

Se desplegará una ventana en la que se introduce información sobre la consulta que queremos analizar. Se introducen las palabras clave y se selecciona el tipo de frase de búsqueda, el motor de búsqueda, el idioma y el número de resultados que se desean analizar:

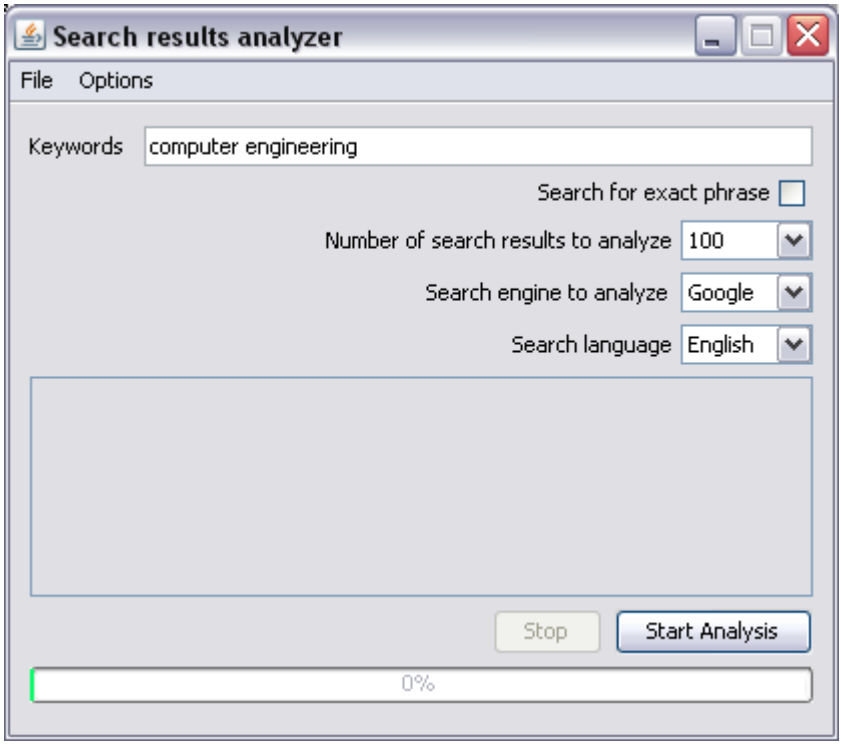


Figura B-11: Introducción en la aplicación de la información relativa a la consulta

Algunos parámetros de ejecución se pueden establecer, también, a fin de reducir el tiempo de ejecución, aunque puede verse comprometida la fiabilidad en la predicción del posicionamiento web. Los valores recomendados por defecto tras los resultados de la experimentación son:

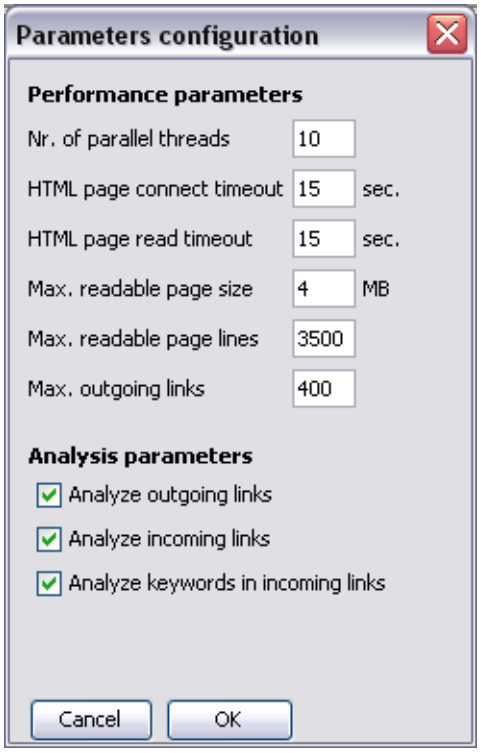


Figura B-12: Parámetros de configuración

Paso 2: Generación de modelos predictivos del posicionamiento web

Después de haber realizado el análisis de resultados de búsqueda para las palabras clave introducidas en la aplicación, se pueden crear modelos capaces de estimar la posición de una página web. La opción del menú principal "Análisis with Weka" permite crear modelos con los algoritmos descritos en esta investigación. Existen en la herramienta otras opciones preparadas para trabajar con otro tipo de algoritmos de aprendizaje automático y así abordar algunos de los desafíos expuestos en el capítulo de trabajos futuros.

Si se desea seleccionar un subconjunto de variables SEO para la construcción de los estimadores, por ejemplo para experimentar con métodos de selección de atributos, existe una opción de activación o desactivación de variables:

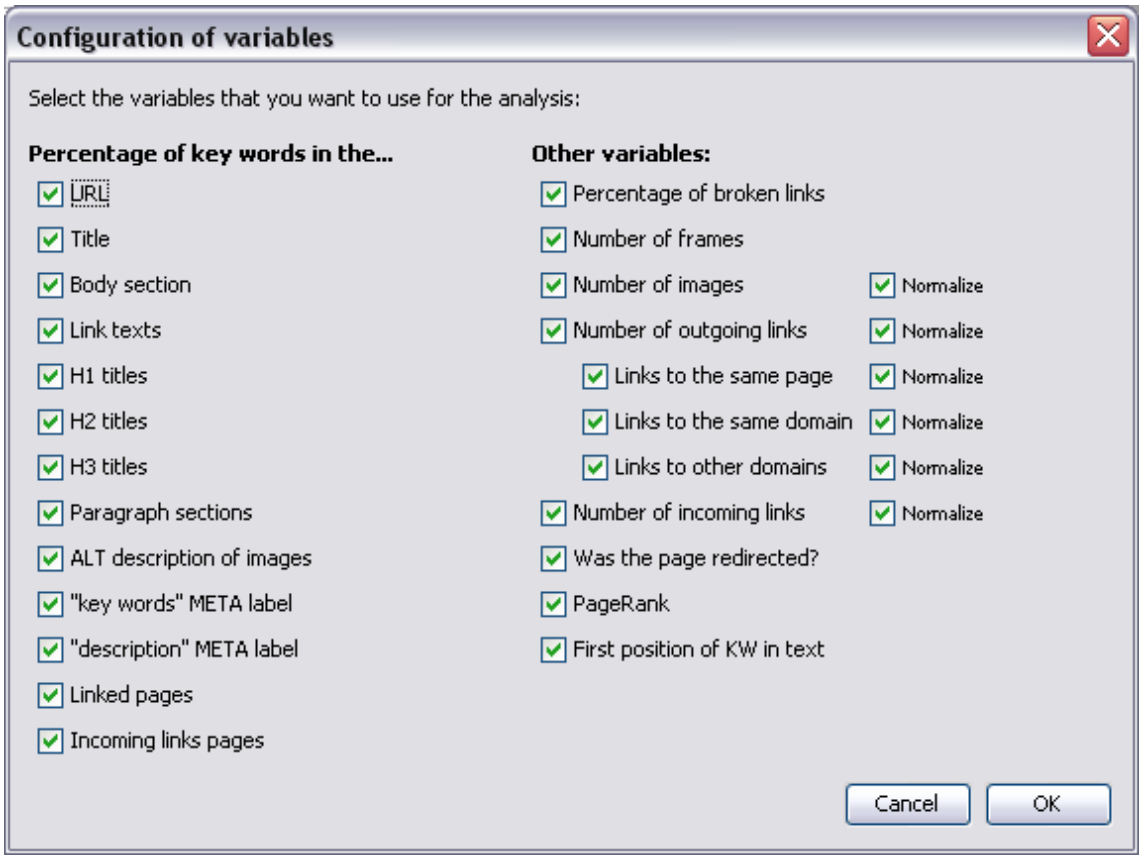


Figura B-13: SEO interfaz de selección de variables

A continuación, se selecciona el motor de búsqueda y uno de los seis algoritmos descritos en la experimentación (todos presentes en la herramienta Weka).

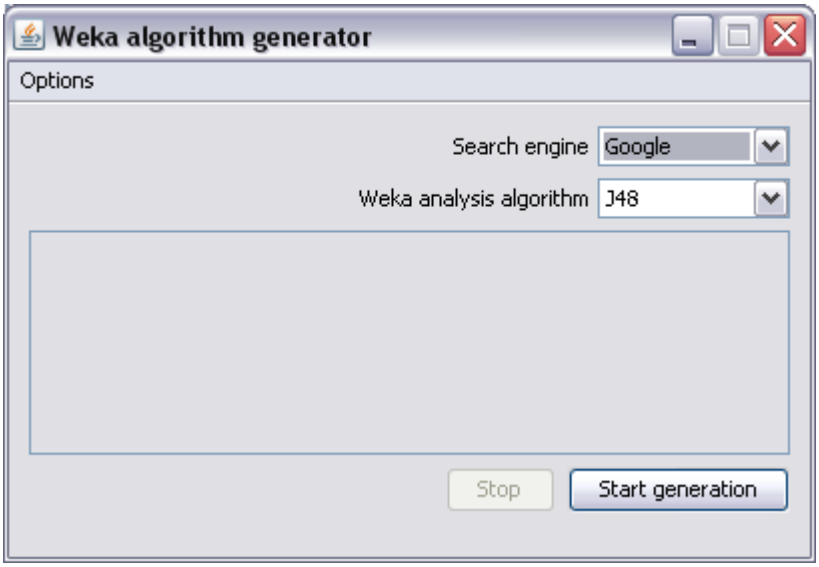


Figura B-14: Selección del buscador de web y del algoritmo de aprendizaje

Una vez seleccionadas las variables SEO que van a intervenir en la generación del modelo (motor de búsqueda y el algoritmo de aprendizaje) se indica a la aplicación, como datos de entrada, el conjunto o conjuntos de resultados de búsqueda procedentes de las consultas efectuadas en el paso anterior. En el siguiente ejemplo, se seleccionan los resultados asociados a las palabras clave "computer engineering" y "genetic algorithms":

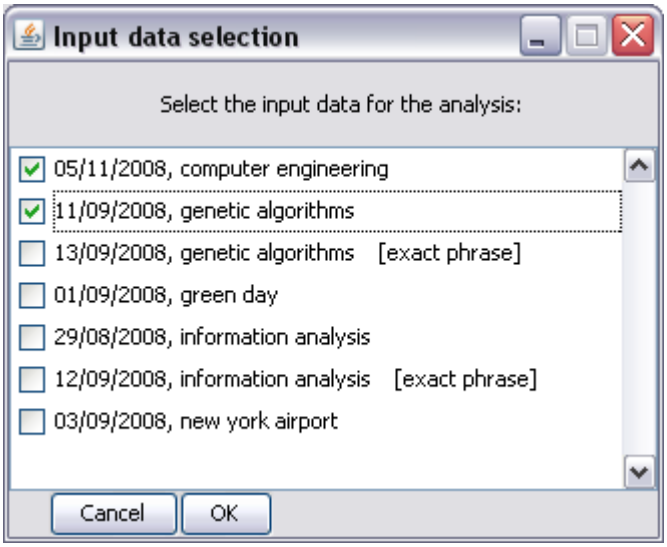


Figura B-15: Selección de los datos de entrada para la creación de modelos

Concluida la construcción del modelo, el programa mostrará en una ventana las tasas de éxito del modelo de estimación generado. El modelo, junto con su índice de éxito, se almacena en la base de datos para uso futuro.

Paso 3: Estimación del posicionamiento

Por último, se puede pasar una página web al programa de modo que pueda predecir cuál es la posición que tendría en el motor de búsqueda seleccionado, para las palabras clave de búsqueda elegidas La siguiente interfaz permite introducir la página web haciendo distinción entre una página cargada mediante dirección web y un archivo HTML local. En este último caso hay que especificar los valores supuestos de las variables (URL, enlaces entrantes, PageRank ...). Se debe concretar también el idioma de la consulta y el modelo en el que se va a basar la predicción. En el caso de que existan varios modelos aplicables para los datos introducidos se puede elegir el más conveniente atendiendo a sus valores de éxito estimados.

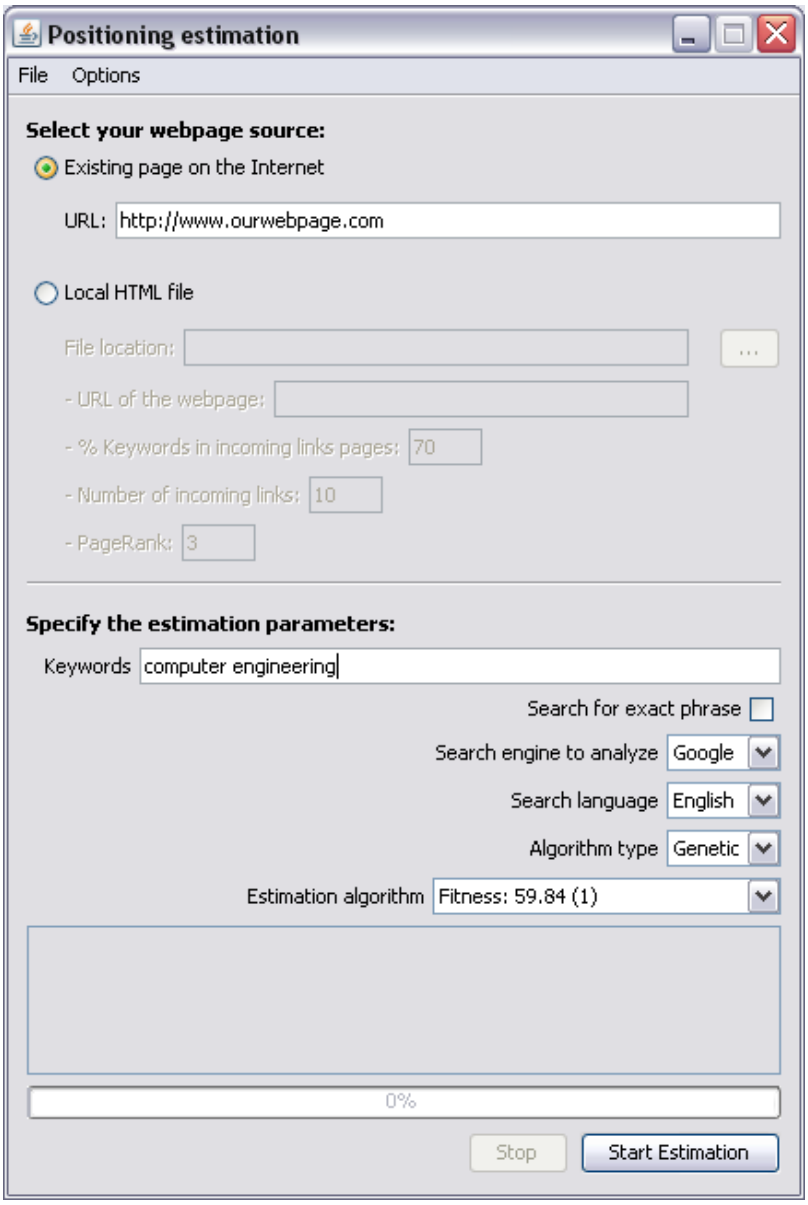


Figura B-16: Interfaz de estimación del posicionamiento de una web

El programa analizará los parámetros SEO de la página web, y aplicará el estimador seleccionado para predecir la posición, mostrando los resultados en una ventana emergente:

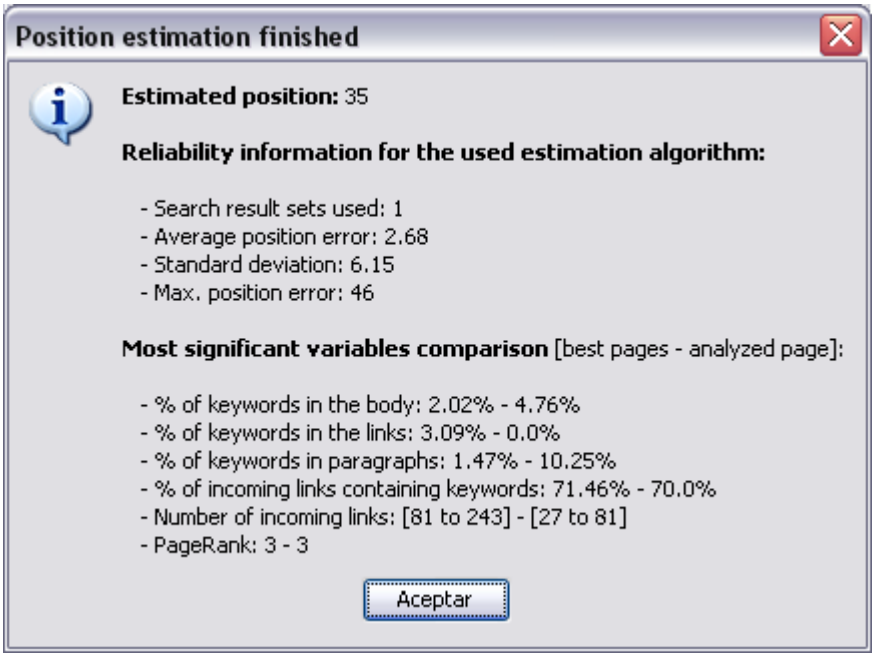


Figura B-17: Resultado del pronóstico de la estimación

Además de informar sobre la tasa de éxito del algoritmo utilizado en la estimación, la herramienta presenta los rangos de los valores de las variables SEO más significativas para los diez mejores resultados de la consulta. La idea es aportar información sobre los rivales más directos en el posicionamiento para esa consulta.

Anexo C: Acrónimos

Abreviatura	Término desarrollado
API	Application Programming Interface
ASCII	American Standard Code for Information Interchange
ARFF	Attribute-Relation File Format
DMOZ	Directory Mozilla
FTP	File Transfer Protocol
GNU	General Public License
GUI	Graphical User Interface
HITS	Hypertext Induced Topic Search
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
HTTPS	HyperText Transfer Protocol Secure
IDF	Inverted Document Frequency
IP	Internet Protocol
IR	Information Retrieval
JAR	Java Archive
JDK	Java Development Kit

JFC	Java Foundation Classes
PDF	Portable Document Format
PR	PageRank
SEO	Search Engine Optimization
TF	Term Frequency
UC3M	Universidad Carlos III de Madrid
UML	Unified Modeling Language
URL	Uniform Resource Locator

Tabla C-1: Acrónimos utilizados